View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by UPCommons. Portal del coneixement obert de la UPC

MÁSTER EN INGENIERÍA DE LOS RECURSOS NATURALES

TRABAJO FINAL DE MÁSTER

ESTUDIO COMPARATIVO SOBRE ACCIDENTES LABORALES EN EL SECTOR MINERO ESPAÑOL, A CIELO ABIERTO Y SUBTERRÁNEO, MEDIANTE EL USO DE TÉCNICAS DE MINERÍA DE DATOS

Autor: Marcos Aznar Cuadrado

Tutor: Josep Maria Rossell

Profesor colaborador: Lluís Sanmiquel

Manresa, Enero 2016

Índice

1. Introducción

1.1. Antecedentes

La seguridad y salud en un ambiente de trabajo es un aspecto de vital importancia tanto para los trabajadores como para las empresas e instituciones responsables de garantizar la salud de los mismos. Teóricamente, se trata de un derecho humano fundamental que debería garantizar la vuelta del trabajador a casa sano y salvo. En este sentido, nadie debería morir o sufrir un accidente en su lugar habitual de trabajo. Sin embargo, en la práctica los registros de la Unión Europea estiman que en la última década murieron aproximadamente un total de 5.500 trabajadores y que, más de 75.000 resultaron gravemente heridos hasta el punto de no poder retomar su trabajo habitual de nuevo (Sanmiquel, et al., 2010).

A pesar de que el gobierno Español haya adoptado nuevas acciones aprobando leyes para reducir el elevado número de muertes y lesiones en el ámbito laboral, España presenta el segundo índice de mortalidad más elevado dentro de la Unión Europea después de Portugal. Uno de los motivos por los que se le atribuye este índice es debido a la tipología de empresas que operan a nivel nacional en este sector. En este sentido, varios estudios (Saari, J. 2005),(Sanmiquel, et al., 2012) han demostrado que la tendencia a que se produzca un accidentes es mayor cuando la empresa es pequeña. Así pues, en España cada día mueren 3 trabajadores y se producen 3.000 lesiones que se traducen en 20 millones de días de bajas anuales (Sanmiquel, et al., 2010). De acuerdo con los datos recogidos por el Ministerio de Empleo y Seguridad Social Español, el sector minero presenta un índice de accidentes por cada 100.000 trabajadores, 4,8 veces mayor que el total de todos los sectores económicos. Este valor todavía es más alarmante si se compara con el índice de países como Estados Unidos y Australia, donde el índice es 7,0 y 6,5 veces mayor, respectivamente (Sanmiquel, et al., 2015).

En general, diversos estudios han comprobado que en aquellas empresas que presentan una peor puntuación en relación a la gestión de la calidad de la seguridad, producen un mayor número de accidentes. Así pues, múltiples autores han publicado trabajos en los que se intenta implementar nuevos y mejores programas de gestión de la seguridad como principal herramienta preventiva (Sanmiquel, et al., 2014). Otros estudios, han determinado que las causas inmediatas del origen del accidente son debidas a: a) las insuficientes condiciones de las instalaciones y las inadecuada maquinaria presente en los lugares habituales de trabajo, b) el comportamiento de los trabajadores y c) las condiciones ambientales del lugar de trabajo (polvo, humedad, desprendimiento de rocas) (Williamson, et al., 1998), (Sanmiquel et al., 2010).

Así pues, el sector de la minería española es sin lugar a duda una de las actividades más peligrosas desarrolladas a nivel nacional. De entre los estudios propuestos, es el de (Sanmiquel et al., 2015) el que determina que las variables *Previous Cases*, *Place*, *Size*, *Physical Activity*, *Preventive Organization*, *Experience* y *Age* forman parte de la génesis de la mayoría de accidentes analizados durante el 2003 y 2012. De entre todas ellas, este trabajo se centra en estudiar las variables y patrones de comportamiento que caracteriza a las actividades desarrolladas a cielo abierto y la subterránea, es decir, en función de su emplazamiento *(Place)*. Otro de los aspectos más interesantes a tener en cuenta, es diferenciar entre los accidentes que tuvieron lugar con y sin sobreesfuerzos. Esta última discriminación es muy importante habida cuenta que los accidentes con sobreesfuerzo están considerados como los más graves dentro del sector de referencia. El objetivo por tanto, no es otro que el de extraer la subyacente información contenida en los datos mediante las técnicas de minería de datos, con la finalidad de mejorar las políticas de prevención y conseguir con ello una reducción de los riesgos, lesiones y muertes en el sector de la minería.

.

1.2. Objetivos

Este trabajo se plantea como objetivo general realizar un estudio comparativo entre: a) los accidentes que tuvieron lugar a cielo abierto y subterráneo y, b) los que tuvieron con y sin sobreesfuerzo. La metodología de trabajo será mediante la combinación de ambas posibilidades, dando lugar así a 4 escenarios diferentes de trabajo.

En definitiva, se trata de encontrar los patrones de comportamiento que caracterizan a cada situación. Es decir, conocer y valorar la importancia de las variables más significativas de cada caso y, observar que importancia tienen y que relación mantienen con la variable respuesta.

Una vez conocidas las relaciones, la información contenida en las mismas puede ser muy útil para la toma de decisiones estratégicas que favorezcan las políticas de prevención y, en consecuencia, contribuyan a la mejora de la seguridad y salud de los trabajadores en el sector de la minería.

A nivel personal, este estudio también pretende introducir al alumno en materia de minería de datos, pues se trata de una reciente y potente técnica mediante la cual se puede articular y extraer una valiosa parte de la información subyacente sobre amplias bases de datos.

Finalmente, comentar que la viabilidad del estudio se ha valorado desde el principio como un punto positivo. Es decir, el trabajo cuenta con el apoyo y la experiencia de dos de los profesores expertos en esta materia.

2. Base de datos y *software*

2.1. Recolección de los datos

Los datos de partida objeto de este estudio fueron proporcionados por el profesor Lluís Sanmiquel, experto en materia de accidentes laborales en el sector minero español. A día hoy trabaja en la Universidad Técnica Superior de Manresa y cuenta con diversas publicaciones referentes para el desarrollo del presente trabajo.

La última publicación (Sanmiquel, et al., 2015) puntualiza que los datos fueron obtenidos de la base de datos digital anual sobre accidentes del Ministerio de Empleo y Seguridad Social de España. Los accidentes considerados en el estudio son los que tuvieron lugar en el sector del trabajo minero durante el plazo ordinario de horas de trabajo y causando a los empleados lesionados la pérdida de, al menos, un día de trabajo. Todo en cuanto se refiere a aquéllos que ocurrieron durante el itinerario de camino de ida o de vuelta al trabajo, quedan excluidos de la base datos.

En consonancia con lo expuesto en dicha publicación, el formato inicial de los ficheros de trabajo fue "*xlx"*, es decir, en formato Excel y en dos ficheros por separado. En el primero se almacenaban todos los accidentes registrados sin sobreesfuerzo y, en el segundo, los que sí se produjeron con sobreesfuerzo. A partir de estos ficheros "madre" se han realizado todas las modificaciones necesarias para desarrollar el estudio. Los programas informáticos con los que se ha elaborado este trabajo son Weka para la minería de datos y Minitab para la estadística descriptiva.

2.2. Estudio de la población por escenarios

La base de datos "madre" en la que se basa el presente estudio está compuesta por un total de 72.250 casos de accidentes de trabajo registrados en el sector de la minería española durante el periodo 2003 y 2013. De este total, 58.345 sucesos corresponden a accidentes sin sobreesfuerzo y, 14.005 a accidentes con sobreesfuerzos. Sin embargo, el estudio comparativo se ha llevado a cabo diferenciando entre cuatro escenarios en los que se discrimina:: a) el lugar del accidente y, b) la presencia o ausencia de sobreesfuerzo en el mismo.

A continuación, se procede a describir las condiciones de los cuatro escenarios considerados durante toda la elaboración del estudio:

- **Escenario 1**

Tabla 1: condiciones y número de casos que definen al escenario 1 (elaboración propia).

Escenario 1		
Emplazamiento *(Place)*	**Sobreesfuerzo**	**nº de casos registrados**
Cielo Abierto (P=3)	No	8427

- **Escenario 2**

Tabla 2: condiciones y número de casos que definen al escenario 2 (elaboración propia). .

Escenario 2		
Emplazamiento *(Place)*	**Sobreesfuerzo**	**nº de casos registrados**
Minería Subterránea (P=4)	No	30042

- **Escenario 3**

Tabla 3: condiciones y número de casos que definen al escenario 3 (elaboración propia).

Escenario 3		
Emplazamiento *(Place)*	**Sobreesfuerzo**	**nº de casos registrados**
Cielo Abierto (P=3)	Sí	2113

- **Escenario 4**

Tabla 4: condiciones y número de casos que definen al escenario 4 (elaboración propia).

Escenario 4		
Emplazamiento *(Place)*	**Sobreesfuerzo**	**nº de casos registrados**
Cielo Abierto (P=4)	Sí	6658

Como se puede apreciar, la suma de datos de los cuatro escenarios, no alcanza el número de casos registrados en la base de datos "madre", pues se han excluido todos los sucesos en los que la variable *Place (P)* no se correspondía a un emplazamiento de tipo cielo abierto (P=3) o de tipo minería subterránea (P=4). En otras palabras, se han eliminado todos los casos registrados de tipo

P= 1, 2 y 5 que correspondían a áreas de almacenamiento, zonas de demolición y construcción y, otros lugares, respectivamente.

Llama especialmente la atención el elevado número de accidentes registrados en el escenario 2, representando el 63,59% de los accidentes. Para el resto de escenarios 1, 3 y 4 esta cifra representa un 17,84%, 4,47% y 14,0% de los datos, respectivamente.

De acuerdo con los escenarios propuestos, se observa que el número de casos total en minería a cielo abierto (escenario 1 y 3) es de 10.540 y para el caso de minería subterránea (escenario 2 y 4) es de 36.700. Así mismo el total de sucesos registrados es de 8.771 cuando se trata de accidentes con sobreesfuerzos (escenario 3 y 4) y de 38.469 (escenario 1 y 2) cuando se estudian los que tuvieron lugar sin sobreesfuerzo. Dicho lo cual, es lógico que el número de casos que contenga nuestra nueva base de datos haya descendido hasta los 47.240 registros, un 34,62% menos de datos respecto a los ficheros "madre".

2.3. Definición de las variables de trabajo

A pesar de que la base de datos inicial de este trabajo recogiera un total de 58 variables, sólo se han considerado las 13 más importantes. El criterio de selección se ha realizado en base a los estudios previos en esta materia, conservando así la nomenclatura, codificación y agrupación establecida en (Sanmiquel et al., 2015). A diferencia de este último estudio, tanto el número de variables respuesta como el de predictoras se han modificado de dos a una y de trece a doce, respectivamente.

Para nuestro estudio, conviene subrayar la exclusión de las variables *Place (P)* y *Lost Working Days (LWD)*. Con respecto a la primera, mencionar que los cuatro escenarios analizados ya contemplan una discriminación explícita del lugar donde se produjo el accidente y que, a su vez, también dependen de si tuvo lugar o no un sobreesfuerzo previo durante el mismo. En cuanto a la segunda, destacar que no es objeto de este estudio valorar la severidad de los accidentes sino que, más bien se centra en:

a) examinar cuáles son las variables más importantes de cada escenario propuesto y,

b) extraer patrones de comportamiento que caractericen a cada tipo de actividad minera comparando los resultados obtenidos.

Por consiguiente, las 12 variables predictoras seleccionadas son las siguientes:

1) **Age (A):** edad (en años) del trabajador lesionado en el momento en el que el accidente sucedió. Se consideran siete franjas de edad para esta variable:

   Clase 1: [16,24]
   Clase 2: [25,29]
   Clase 3: [30,34]
   Clase 4: [35,39]
   Clase 5: [40, 44]
   Clase 6: [45,54]
   Clase 7: [55 o más]

2) **Experience (E):** experiencia (en meses) del trabajador en el tipo de cargo y tareas que le son encomendadas. Se han considerado los siguientes siete grupos:

Clase 1: [0,12]
Clase 2: [13,30]
Clase 3: [31,60]
Clase 4: [61,120]
Clase 5: [121,180]
Clase 6: [181, 240]
Clase 7: [241 o más]

3) **Size (S):** tamaño de la empresa en función del número de trabajadores que están dados de alta en la compañía. La clasificación se divide en seis grupos:

Clase 1: [0,9]
Clase 2: [10, l9]
Clase 3: [20,49]
Clase 4: [50,99]
Clase 5: [100, 499]
Clase 6: [500 o más]

4) **Contract (C):** tipología de contrato laboral dividido en cuatro clases. La clasificación se realiza en función de la temporalidad y el número de horas de la jornada laboral. Así pues, diferenciamos entre cuatro clases de contratación:

Clase 1: contratación indefinida a jornada completa
Clase 2: contratación indefinida a jornada parcial
Clase 3: contratación temporal a jornada completa
Clase 4: contratación temporal a jornada parcial

5) **Previous Causes (PC):** existencia de causas previas antes de que tuviera lugar el accidente. Todas ellas se han codificado en siete categorías:

Clase 1: problema eléctrico, explosión, fuga, desbordamiento, derrame, vaporización, vuelco
Clase 2: rotura, fractura, estallido, resbalón, caída, derrumbamiento de agente material
Clase 3: pérdida de control (total o parcial) de control de máquinas
Clase 4: caídas de personas
Clase 5: movimiento del cuerpo sin esfuerzo físico
Clase 6: movimiento del cuerpo con esfuerzo físico
Clase 7: otras

6) **Physical Activity (PA):** tipo de actividad física que el trabajador estaba desarrollando en el momento en el que tuvo lugar el accidente. Éstas han sido agrupadas mediante el uso de siete categorías:

Clase 1: operaciones con máquinas
Clase 2: trabajos con herramientas manuales
Clase 3: conducir o estar a bordo de un medio de transporte
Clase 4: manipulación de objetos
Clase 5: transporte manual de cargas
Clase 6: realización de un movimiento
Clase 7: otras

7) **Preventive Organization (PO):** servicio y organización de los recursos necesarios para el desarrollo de las actividades preventivas. En este caso, el empresario dispone de seis opciones para afrontar su deber de prevención:

Clase 1: asunción personal por el propio empresario.
Clase 2: designación de trabajadores formados que se encargarán de dicha prevención.
Clase 3: servicio de prevención propio.
Clase 4: servicio de prevención mancomunado
Clase 5: servicio de prevención externo.
Clase 6: sin servicio de prevención.

8) **Risk (R):** variable en la que se determina la toma de conciencia de los riesgos que supone la actividad por parte de la empresa, es decir, se determina si la compañía realizó o no estudios previos al respecto. Se diferencian dos clases:

Clase 0: no hubo evaluación previa
Clase 1: sí hubo evaluación previa

9) **Day Week (DW):** día de la semana en el que tuvo lugar el accidente y siete obvias clases a tener en cuenta:

Clase 1: lunes
Clase 2: martes
Clase 3: miércoles
Clase 4: jueves
Clase 5: viernes
Clase 6: sábado
Clase 7: domingo

10) **Hour Day (HD):** hora del día (en sistema 24h) en el que se registró el accidente laboral. Se agrupan las veinte y cuatro horas del día en cinco categorías:

Clase 1: (0,6]
Clase 2: (6,10]
Clase 3: (10,14]
Clase 4: (14,18]
Clase 5: (18, 24]

11) **Work Hours (WH):** número de horas trabajadas por parte del empleado justo antes de que tuviera lugar el accidente. Se han establecido seis franjas horarias:

Clase 1: (0, 1]
Clase 2: (1, 4]
Clase 3: (4, 8]
Clase 4: (8, 10]
Clase 5: (10, 12]
Clase 6: (12 o más]

12) **Contractual Status (CS):** tipo de relación que tiene el trabajador con la empresa, es decir, si la contratación es a través de una empresa externa o no. Así, pues se diferencian dos tipos de relaciones contractuales:

Clase 0: la empresa que contrata es subcontratada
Clase 1: la empresa que contrata es la contratista

La variable respuesta de este estudio es la siguiente:

13) **Type of Accident (TA):** causa del accidente, es decir, el evento en cuestión que produjo la lesión al trabajador. En este sentido, se dividen en seis categorías:
Clase 1: contacto eléctrico, fuego, contacto con sustancias peligrosas. Ahogamiento, quedar sepultado o envuelto
Clase 2: golpe contra un objeto estacionario
Clase 3: choque o golpe contra un objeto en movimiento o colisión
Clase 4: contacto con objeto cortante, punzante, duro o rugoso
Clase 5: sobreesfuerzo físico, trauma psíquico, radiaciones, ruido, luz o presión
Clase 6: otras

Finalmente, cabe señalar que todas estas variables serán objeto del posterior estudio de minería de datos mediante el software Weka. Este programa nos permitirá extraer información subyacente en relación a los patrones de comportamiento y a las variables más importantes para cada estadio considerado. En cuanto a la estadística descriptiva, es necesario también aclarar que, por consiguiente, se analizarán las variables predictoras más influyentes para cada uno de los escenarios propuestos.

2.4. Programas informáticos utilizados

2.4.1. Waikato Environmental for Knowledge Analysis (Weka)

En el presente estudio se va a trabajar con Weka (*Waikato Environtment for Knowledge Analysis)*, un *software* programado en Java que potencia la extracción de conocimiento a partir de bases de datos con gran cantidad de información. Es original de Nueva Zelanda y se desarrolló en la Universidad de Waikato. Existen otras herramientas similares tales como *Oracle Data Miner* o *Clementine*, pero los motivos por los que se ha decidido trabajar con Weka son principalmente por su licencia GNL *(General Public License)* y por el conocimiento que el tutor de este trabajo tiene sobre este programa en sus previas experiencias y publicaciones. Se trata por tanto, de un *software* con garantías: buen funcionamiento, alta confiabilidad y elevado potencial.

En este sentido, Weka es una colección de algoritmos de aprendizaje automático para tareas relacionadas con la minería de datos. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos (Witten, et al., 2011).

Todas las técnicas disponibles en Weka se realizan sobre datos de entrada que se encuentren codificados en formato *arff (attribute-relation file format)*. Así pues, el *software* permite cargar los datos en tres tipos de soporte: a) fichero de texto, b) acceso a una base de datos y, c) acceso a través de internet sobre una dirección URL de un servidor web. En nuestro estudio, se trabajará con ficheros de texto cuyos datos deben estar dispuestos de la forma siguiente:

```
@relation accidentes_sinsobreesfuerzo_3

@attribute A {"1","2","3","4","5","6","7"}
@attribute E {"1","2","3","4","5","6","7"}
@attribute S {"1","2","3","4","5","6"}
@attribute C {"1","2","3","4"}
@attribute PC {"1","2","3","4","5","6","7"}
@attribute PA {"1","2","3","4","5","6","7"}
@attribute PO {"1","2","3","4","5","6"}
@attribute R {"0","1"}
@attribute DW {"1","2","3","4","5","6","7"}
@attribute HD {"1","2","3","4","5"}
@attribute WH {"1","2","3","4","5","6"}
@attribute CS {"0","1"}
@attribute TA {"1","2","3","4","5","6"}

@data
7,7,2,1,7,6,5,1,4,4,3,1,6
6,7,3,1,5,2,5,1,2,2,1,0,3
6,4,1,1,6,6,5,0,4,3,2,0,4
4,4,5,1,2,6,5,1,6,3,3,0,5
6,5,5,1,7,6,4,1,2,3,3,0,6
5,3,2,3,5,7,5,1,3,3,2,0,4
4,2,2,1,3,1,5,1,4,4,3,0,6
3,2,2,1,6,6,5,1,2,2,1,0,2
7,2,3,3,2,7,5,1,4,3,3,0,3
6,5,1,1,4,6,5,0,5,4,3,0,2
4,5,5,1,4,6,3,1,2,3,3,0,2
```

Figura 1: ejemplo del archivo "escenario 1" en formato csv previamente a ser transformado al formato arff. Elaborado en un bloc de notas (elaboración propia).

Si se quiere agilizar este proceso, se recomienda aprovechar el fichero original con todas las variables de interés en formato Excel. Una vez seleccionadas las variables, se debe guardar el archivo en formatos tipo CSV (delimitado por comas) (*.csv). De este modo, se obtendrán los datos almacenados en columnas y filas separados por comas. Después, se abre dicho archivo mediante el bloc de notas y se realizan las modificaciones que, a continuación, se describen. Es necesario recalcar que, al trabajar con un fichero de texto *txt* (bloc de notas) se debe introducir manualmente ".arff" al final del archivo. Sólo así Weka reconocerá la base de datos con la que se desee trabajar.

Como se puede observar en la *figura 1*, cada fichero debe empezar por "@relation" más el nombre del fichero que se pretenda utilizar, en este ejemplo se han adjuntado una reducida parte de los datos del escenario 1, donde no hay sobreesfuerzo y el emplazamiento es a cielo abierto (P=3). A continuación, se deben añadir los atributos mediante el signo "@attribute" más el nombre de la variable. En nuestro caso, al tratarse de datos categóricos se debe añadir entre llaves todas las clases que pertenecen a dicho atributo. Además, cada una de éstas debe ir entre comillas y separada por una coma. Finalmente y, tras escribir "@data" se añade el conjunto de datos, donde cada fila representa un caso y cada columna una variable. Como ejemplo, se puede observar que la primera fila representa al primer accidente laboral minero registrado para el escenario 1 y que, cada valor numérico se asocia en orden con los atributos descritos previamente. Así pues, se aprecia que en el escenario 1, el primer caso registrado acoge los siguientes valores de atributos:

A=7, E=7, S=2, C=1, PC=7, PA=6, PO=5, R=1, DW=4, HD=4, WH=3, CS=1, TA=6

De la misma manera se interpreta el resto de valores categóricos que aparecen en el fichero de la *Figura 1*.

En cuanto al número de ficheros se refiere, se han construido 4 ficheros base igual al expuesto con las condiciones que caracteriza a cada uno de los escenarios propuestos.

2.4.2. Minitab

Minitab es un programa informático desarrollado en la Universidad del Estado de Pensilvania y está diseñado para ejecutar funciones estadísticas básicas y avanzadas: gestión de datos y archivos, análisis de regresión, potencia y tamaño de la muestra, tablas y gráficos, análisis multivariante, control estadístico de procesos, análisis del sistema de medición, análisis de varianza entre otras.

La estadística descriptiva de este trabajo se ha realizado mediante este *software* habida cuenta que, en la reciente asignatura cursada del máster *"técnicas de análisis estadístico de datos y diseño y planificación de experimentos"* éste fue el programa de referencia. Los fichero de trabajo han sido dos, diferenciando entre accidentes laborales con y sin sobreesfuerzo. Las variables analizadas en cada uno de ellos corresponden a las expuestas en el apartado 2.3 y los emplazamientos contemplados: cielo abierto (P=3) y minería subterránea (P=4).

3. Minería de datos con Weka

3.1. Introducción a la Minería de Datos

La minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacía determinadas técnicas especializadas y englobadas bajo el nombre de minería de datos o *Data Mining.*

Actualmente la cantidad de datos en el mundo y en nuestras vidas parece incrementarse cada vez más indefinidamente. Por un lado, las nuevas tecnologías han facilitado el almacenamiento de datos en todo tipo de medios y aparatos tales como ordenadores, *smart phones*, discos duros y servidores en línea. Ello ha promovido la tendencia a aplazar decisiones sobre el qué hacer con todos estos datos. En consecuencia, nuestras compras en los supermercados, nuestros hábitos financieros y nuestro ir y venir acaban quedando registrados en bases de datos. Por el otro lado, la aparición de internet y su consolidación dentro de nuestra sociedad, ha transformado radicalmente la forma que teníamos de entender y conseguir la información. Así pues, cada elección que tomamos acaba siendo almacenada. A medida que el volumen de datos vaya aumentado inexorablemente, la proporción de lo que la gente entienda de éstos va disminuyendo de una forma alarmante (Witten, et al., 2011). Es obvio por tanto, que la información subyacente de estos datos se defina como potencialmente útil si bien se sabe articular e interpretar eficazmente.

En cuanto a las aplicaciones de la minería de datos, llama la atención la diversidad de campos que puede llegar abarcar, siendo el de la ciencia el que a día de hoy, está extrayendo mayor rendimiento. De entre ellos, destacan las aplicaciones en el sector financiero y de la banca, el de análisis de mercados y comercio, el de seguros y salud privada, educación, procesos industriales, medicina, biología y bioingeniería, telecomunicaciones y muchas otras áreas. Uno de los factores más relevantes de la aplicación de la minería de datos, es la comprensión y el conocimiento de los propios conceptos que la definen, independientemente de cual sea el campo para el que se aplique. Así pues, los programas informáticos permiten obtener resultados sin necesidad de descifrar el desarrollo matemático de los algoritmos que se encuentran debajo de los procedimientos.
No obstante, la minería de datos es ya un concepto muy evolucionado que necesita ser aproximado conceptualmente por etapas. Inicialmente la finalidad de los sistemas de información era recopilar

información sobre una parcela determinada para ayudar en la toma de decisiones. Con la informatización de las organizaciones y la aparición de aplicaciones software operacionales sobre el sistema de información, la finalidad principal de los sistemas de información es dar soporte a los procesos básicos de la organización (ventas, producción, personal…). Una vez satisfecha la necesidad de tener un soporte informático para los procesos básicos de la organización, es decir, los llamados "sistemas de información para la gestión", las organizaciones exigen nuevas prestaciones de los sistemas de información. Éstos, reciben del nombre de "sistemas de información para la toma de decisiones".

Como resultado, han aparecidos las DSS *(Decision Support Systems)*, herramientas de negocio para la toma de decisiones. Éstas permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y así predecir comportamientos futuros. La Minería de Datos analiza los datos mientras que las herramientas facilitan el acceso a la información y aumentan la eficacia de su análisis con el fin de potenciar la toma de decisiones estratégicas.

Pero la minería de datos es sólo una de las etapas que se engloba dentro de todo el proceso de extracción de conocimiento a partir de datos, también conocido como KDD (*Knowledge Discovery in Database)*. En este sentido, el proceso consta también de varias fases como son la preparación de datos (selección, limpieza y transformación), su exploración y auditoria, minería de datos propiamente dicha (desarrollo de modelos y análisis de datos), evaluación, difusión y utilización de modelos (*output).* Además, el proceso de extracción del conocimiento incorpora muy diferente técnicas (árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas bayesianas, máquinas de soporte vectorial, etc) de campos diversos (aprendizaje automático e inteligencia artificial), estadística, bases de datos y aborda un tipología variada de problemas (clasificación, categorización, estimación/regresión, agrupamiento, etc.) (Pérez, et al., 2007).

Por lo tanto, se puede deducir que durante todo el proceso de extracción del conocimiento (KDD) se establece una serie de fases que se enumeran a continuación:

1. Selección
2. Exploración
3. Limpieza
4. Transformación
5. Minería de Datos
6. Evaluación
7. Difusión

Las cuatro primeras fases se suelen englobar bajo el nombre de preparación de datos. Durante la fase de selección se recopilan los datos, se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas, se identifican y seleccionan las variables relevantes en los datos y se aplican las técnicas de muestreo adecuadas. Debido a que los datos pueden provenir de diferentes fuentes, es necesario una etapa de exploración mediante técnicas de exploratorio de datos, buscando entre otras cosas la distribución de los datos, su simetría y normalidad y las correlaciones existentes en la información. A continuación es necesaria la limpieza de los datos, ya que pueden contener valores atípicos, valores faltantes y valores erróneos. En esta fase se analiza la influencia de los datos atípicos, se imputan los valores faltantes y se eliminan o corrigen los datos incorrectos. A continuación, si es necesario, se lleva a cabo la transformación de los datos, generalmente mediante técnicas de reducción o aumento de la dimensión y escalado simple y multidimensional, entre otras.
En la fase de minería de datos, se decide cuál es la tarea a realizar (clasificar, agrupar, etc) y se elige la técnica descriptiva o predictiva que se va a utilizar. En la fase de evaluación e interpretación se evalúan los patrones y se analizan por expertos, y si es necesario se vuelve a las

fases anteriores para una nueva iteración. Finalmente, en la fase de difusión se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios.

En cuanto a las técnicas de minería de datos se distingue entre técnicas predictivas, en las que las variables pueden clasificarse inicialmente en dependientes o independiente, técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus y técnicas auxiliares.

En las primeras, se especifica el modelo para los datos en base a un conocimiento teórico previo. Las redes neuronales permiten descubrir modelos complejos y afinarlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Se pueden incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminantes, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas.

En lo que se refiere a técnicas descriptivas, no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de *clustering* y segmentación, las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión y de escalamiento multidimensional.

Tanto las primeras como las segundas, se centran en el descubrimiento del conocimiento embebido de los datos.

Las técnicas auxiliares son herramientas de apoyo más superficiales y limitadas. Se trata de nuevos métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la verificación.

Como ya se ha mencionado anteriormente, en la elaboración de este estudio comparativo se va a trabajar la minería de datos con el *software* informático Weka. Un programa ampliamente utilizado en esta materia que, a día de hoy, cuenta con un amplio abanico de trabajos y publicaciones en diferentes campos que lo respaldan (Sanmiquel, et al.,2015).

3.2. Metodología de trabajo y procedimientos estadísticos mediante Weka

La metodología de trabajo desarrollada en el presente estudio repite parte de las pautas establecidas en (Sanmiquel, et al., 2015). Así pues, los procedimientos estadísticos que se han aplicado son los siguientes:

3.2.1. Selección de variables significativas

El motivo por el cual se realiza esta selección, es debido al elevado número de casos (72.250 accidentes) y variables (58) que registraba la base de datos inicial. En el caso de trabajar con ésta, se dificultaría el proceso de extracción de información en términos de tiempo y calidad. Es por tanto lógico que, toda esta información se articulará de acuerdo con los objetivos prefijados en la elaboración de este estudio.

Dicho lo anterior, se han seleccionado un total de 13 variables según los criterios expuestos en el apartado 2.3. Por consiguiente, el número de accidentes se ha visto reducido un 35%, alcanzando así los 47.240 sucesos.

3.2.2. Comprobación de la fiabilidad de los datos

Se realiza un análisis de fiabilidad de datos para cada variable seleccionada. Los procedimientos utilizados es esta fase de comprobación son gráficos y estadísticos numéricos. Los primeros, se refieren a gráficos de barras de todos los atributos; así se verifica que el número de barras se corresponde con el de los diferentes grupos clasificados inicialmente. Los segundos, incluyen el cálculo de la media, mediana, desviación estándar, el mínimo y el máximo. Todos ellos se recopilan en el apartado 4.2 de este trabajo.

Una vez se han elaborado ambos, es factible encontrar los posibles valores atípicos *(outliers)* o errores en la clasificación. Los datos erróneos son objeto de una corrección o supresión según el suceso del cual se trate. En nuestra base de datos únicamente se encontraron desviaciones en relación al número de casos (N) que almacenaba cada variable. Es decir, los datos de algunas variables no se correspondían con el número total de accidentes registrados, sin embargo el problema fue resuelto.

3.2.3. Codificación de las variables

Las variables seleccionadas se agrupan en clases al objeto de tratarlas como variables categóricas, es decir, cada unidad de observación debe catalogarse sin ambigüedad en una y sólo una de las categorías posibles. Además, se debe poder clasificar todo suceso sin excepción alguna.

Para el presente estudio, la codificación de las variables se ha respetado según lo expuesto en el apartado 2.3, donde se matiza que se ha conservado la codificación de trabajos previos. Únicamente se han modificado categorías de variables en las que el número de clases de la base de datos no se correspondía con el del estudio de (Sanmiquel, et al., 2015).

3.2.4. Aplicación de clasificadores Weka

Al objeto de seleccionar las variables predicotras más importantes, se han aplicado varios métodos de clasificación mediante la opción *Select Attributes*. Esta opción permite automatizar la búsqueda de los subconjuntos más apropiados para la variable de salida. Todos ellos, se recopilan en la t*abla 5*.

Tabla 5: esquema de la metodología aplicada al estudio: *Attribute evaluators, Search methods* y *Attribute selection modes* (Sanmiquel, et al., 2015).

Response Attribute	Attribute evaluators	Search methods	Attribute selection modes
Type of Accident	OneRAttributeEval	Ranker	3, 5 and 10 cross-validation
	ChiSquaredAttributeEval	Ranker	
	CfsSubsetEval	GreedyStepwise	Full training set
		ExhaustiveSearch	
		BestFirst	
	ClassifierSubsetEval	RandomSearch	
	InfoGainAttributeEval	Ranker	

En el proceso de clasificación se han utilizado 5 evaluadores de atributos combinados con 5 métodos de búsqueda y 2 técnicas estadísticas. La variable respuesta en todos los casos ha sido TA *(Type of Accident)*.

A continuación, se describen brevemente los tipos de evaluadores, métodos de búsqueda y las técnicas estadísticas que se han empleado (Witten, et al., 2011):

- **Attribute evaluators**

  En primer lugar se describirán los evaluadores de atributos individuales *(single-attribute evaluators)*. Éstos toman muestras de datos aleatoriamente y verifica casos parecidos de la misma y diferente clase.

  - ***OneRAttributeEval:*** utiliza la medida de precisión sencilla adoptada por el clasificador *OneR*, el cual usa el atributo de mínimo error para predecir mediante la discretización los atributos numéricos. Además, este clasificador es de los más sencillos y rápidos; sus resultados son muy buenos en comparación con algoritmos mucho más complejos.

  - ***ChiSquaredAttributeEval:*** calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase. De este modo, se obtiene el nivel de correlación entre la clase y cada atributo.

  - ***InfoGainAttributeEval:*** evalúa los atributos midiendo su ganancia de información de cada uno con respecto a la clase.

  En segundo lugar, se explican los evaluadores de subconjuntos, llamados *Subset Evaluators* y cuyo funcionamiento se rige por la selección de subconjuntos de atributos que devuelven una medida numérica que guía la búsqueda.

  - ***CfsSubsetEval:*** evalúa un subconjunto de atributos considerando la habilidad predictiva individual de cada variable, así como el grado de redundancia entre ellas. Se escogen los subconjuntos de atributos que estén altamente correlacionados con la clase y que tengan baja intercorrelación.

  - ***ClassifierSubsetEval:*** evalúa los subconjuntos de atributos en los datos de entrenamiento o en un conjunto de prueba independiente utilizando un clasificador.

- **Search methods**

  Los métodos de búsqueda funcionan atravesando los espacios generados por los atributos para así encontrar un buen subconjunto. Se mide la calidad mediante el evaluador de subconjuntos de atributos seleccionado. Cada método de búsqueda se puede configurar con el editor de Weka.

  - ***BestFirst:*** esta búsqueda se realiza mediante una escalada con retroceso; se puede especificar el número de nudos consecutivos que deben ser encontrados antes de que el sistema dé marcha atrás.

  - ***Ranker*:** es un método de búsqueda que ordena los atributos por sus evaluaciones individuales y por tanto, únicamente se aplica *single-attribute evaluators.* Además, este método también realiza un proceso de selección de atributos mediante la eliminación de aquéllos que ocupan las posiciones del ranking más bajas.
  - ***GreedyStepwise:*** este método realiza la búsqueda con avidez a través del espacio de subconjunto de atributos progresando hacía adelante desde el conjunto vacío o hacía atrás del conjunto completo. En un modo alternativo, ocupa atributos por los que atraviesa el espacio de vacío a lleno (o viceversa) y graba el orden en que

se seleccionan los aquéllos. Se puede especificar el número de atributos para retener o fijar un umbral por debajo del cual los atributos se descartan.

- ***ExhaustiveSearch:*** este método busca en el espacio subconjuntos de atributos a partir de un conjunto vacío y genera los informes del mejor subconjunto encontrado. Si se establece un conjunto inicial, éste buscara hacía atrás desde este punto y generará informes sobre el subconjunto más pequeño (o igual) con una mejor evaluación.

- ***RandomSearch:*** este método busca al azar el espacio de subconjunto de atributos. Si se suministra un conjunto inicial, busca subconjunto que mejoren (o igualen) el del punto de partida y que tenga menos números (o el mismo) de atributos. De lo contrario, empezará desde un punto al azar y se hallarán los resultados de los mejores subconjuntos encontrados.

- **Attribute selection modes**

  - ***Full training set:*** método de entrenamiento que implica usar todo el conjunto de datos para el valor del subconjunto de atributos evaluado.

  - ***n fold cross-validation*****:** los datos se dividen en *n* subconjuntos para realizar *n* iteraciones. Uno de los subconjuntos se utiliza como datos de prueba y el resto (*n-1)* como datos de entrenamiento. El proceso de validación cruzada es repetido durante *n* iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente, se realiza la media aritmética de los resultados de cada iteración y así obtener un único resultado. Se trata de un método preciso habida cuenta que se realiza la evaluación a partir de *n* combinaciones de datos de entrenamiento y prueba. Sin embargo, el inconveniente es que desde el punto de vista computacional, el método resulta más lento aunque más preciso. En la práctica, el número de iteraciones *n* que se escoge depende de la medida del conjunto de datos. Suele ser habitual trabajar la validación cruzada con un máximo de 10 iteraciones. Para nuestro estudio, se ha trabajado con *n*=3, 5 y 10. Este método es muy útil para evaluar modelos en los que se utilizan varios clasificadores (Witten, et al., 2011).

A continuación, se presentan las hojas de cálculo de los cuatro escenarios estudiados. Por un lado, se almacenan los valores de los algoritmos matemáticos definidos en las *tablas 6, 8,10 y 12* y, por el otro se presenta en las *tablas 7, 9,11 y 13* los estadísticos numéricos que se han utilizado para valorar el posicionamiento de las siete variables más significativas.

En cuanto al método de selección, subrayar que el aspecto más importante a tener en cuenta, es el ranking de la variable en cada prueba tipo *single-attribute evaluator*. El valor de los algoritmos acogen valores no equiparables, pues no se rigen bajo las mismas operaciones matemáticas y, por tanto, no tendría ningún sentido realizar ningún tipo de ponderación al respecto. Queda claro por tanto, que el mejor método para determinar las variables más significativas, se rija por los valores del ranking de cada prueba. Para ello, se han tenido en cuenta las veces que se repetía cada valor y se han calculado diversos estadísticos numéricos: la media, la mediana, la moda y la desviación estándar. Respecto a las pruebas *attribute subset evaluators*, destacar que su aportación ha sido más bien de verificación. Se han comprobado que de las pruebas realizadas, la mayoría de las siete variables significativas fuesen escogidas como mínimo en un tipo de prueba.

Escenario 1

Tabla 6: resultados la clasificación de las variables más significativas para el escenario 1 (elaboración propia).

Attribute Evaluator	Search Method	Attribute selection modes	A	E	S	C	PC	PA	PO	R	DW	HD	WH	CS
Single-Attribute Evaluators														
OneRAttributeEval	**Ranker**	Full Training set	30.889	30.972	30.936	30.758	53.803	42.874	31.684	30.948	30.343	30.497	30.746	30.533
		ranking	**7**	**4**	**6**	**8**	**1**	**2**	**3**	**5**	**12**	**11**	**9**	**10**
		Cross-validation (n=3)	30.752	30.764	30.592	30.877	53.803	42.744	31.524	30.883	30.361	30.675	30.574	30.355
		ranking	**6**	**7**	**9**	**4**	**1**	**2**	**3**	**5**	**12**	**8**	**10**	**11**
		Cross-validation (n=5)	30.841	30.859	30.732	30.568	53.803	42.830	31.660	30.791	30.254	30.242	30.515	30.405
		ranking	**5**	**7**	**6**	**8**	**1**	**2**	**3**	**4**	**12**	**11**	**9**	**10**
		Cross-validation (n=10)	30.745	30.868	30.848	30.549	53.803	42.832	31.660	30.923	30.164	30.434	30.616	30.467
		ranking	**7**	**6**	**5**	**9**	**1**	**2**	**3**	**4**	**12**	**11**	**8**	**10**
ChiSquaredAttributeEval	**Ranker**	Full Training set	62.239	90.536	127.275	62.778	7294.954	1925.721	127.444	15.222	48.256	33.619	30.389	5.757
		ranking	**7**	**5**	**4**	**6**	**1**	**2**	**3**	**11**	**8**	**9**	**10**	**12**
		Cross-validation (n=3)	56.359	69.925	91.284	46.397	4869.654	1292.659	98.440	11.656	40.955	28.375	30.345	9.516
		ranking	**6**	**5**	**4**	**7**	**1**	**2**	**3**	**11**	**8**	**10**	**9**	**12**
		Cross-validation (n=5)	55.630	77.870	107.205	52.589	5841.371	1548.323	107.963	13.136	46.580	30.162	28.167	5.508
		ranking	**6**	**5**	**3**	**7**	**1**	**2**	**4**	**11**	**8**	**9**	**10**	**12**

		Cross-validation (n=10)	59.490	84.167	1170.049	57.663	6568.307	1737.001	117.191	14.190	47.288	31.815	29.579	5.649
		ranking	**6**	**5**	**4**	**7**	**1**	**2**	**3**	**11**	**8**	**9**	**10**	**12**
InfoGainAttributeEval	**Ranker**	Full Training set	0.005464	0.007728	0.010613	0.0058	0.465233	0.163596	0.010112	0.001294	0.003954	0.002918	0.002693	0.000476
		ranking	**7**	**5**	**3**	**6**	**1**	**2**	**4**	**11**	**8**	**9**	**10**	**12**
		Cross-validation (n=3)	0.007	0.009	0.012	0.006	0.466	0.165	0.012	0.001	0.005	0.004	0.004	0.001
		ranking	**6**	**5**	**3**	**7**	**1**	**2**	**4**	**11**	**8**	**10**	**9**	**12**
		Cross-validation (n=5)	0.006	0.008	0.011	0.006	0.466	0.164	0.011	0.001	0.005	0.003	0.003	0.001
		ranking	**6**	**5**	**3**	**7**	**1**	**2**	**4**	**11**	**8**	**9**	**10**	**12**
		Cross-validation (n=10)	0.006	0.008	0.011	0.006	0.466	0.164	0.01	0.001	0.004	0.003	0.003	0.001
		ranking	**7**	**5**	**4**	**6**	**1**	**2**	**3**	**11**	**8**	**9**	**10**	**12**
Attribute Subset Evaluators														
CfsSubseEval	**GreedyStepwise**	Full Training set	×	×	×	×	✓	×	✓	×	×	×	×	×
		Cross-validation (n=3)	0	0	0	0	3	0	3	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	1	0	5	0	4	0	0	0	0	0
		number of folds (%)	**0**	**0**	**20**	**0**	**100**	**0**	**80**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	0	0	0	0	10	0	10	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**

	Exhaustive Search	Full Training set	×	×	×	×	✓	×	✓	×	×	×	×	×
		Cross-validation (n=3)	0	0	0	0	3	0	3	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	1	0	5	0	4	0	0	0	0	0
		number of folds (%)	**0**	**0**	**20**	**0**	**100**	**0**	**80**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	0	0	0	0	10	0	10	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**
	BestFirst	Full Training set	×	×	×	×	✓	×	✓	×	×	×	×	×
		Cross-validation (n=3)	0	0	0	0	3	0	3	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	1	0	5	0	4	0	0	0	0	0
		number of folds (%)	**0**	**0**	**20**	**0**	**100**	**0**	**80**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	0	0	0	0	10	0	10	0	0	0	0	0
		number of folds (%)	**0**	**0**	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**0**
ClassifierSubsetEval	**RandomSearch**	Full Training set	×	✓	✓	✓	×	✓	×	×	×	×	✓	✓
		Cross-validation (n=3)	0	3	3	3	0	3	0	0	0	0	3	3

	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=5)	0	5	5	5	0	5	0	0	0	0	5	5
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=10)	0	10	10	10	0	10	0	0	0	0	10	10
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**

Tabla 7: resultados de los estadísticos numéricos calculados a partir de la *Tabla 6.* Selección de las 7 variables más significativas para el escenario 1 (elaboración propia).

Variables	**A**	**E**	**S**	**C**	**PC**	**PA**	**PO**	**R**	**DW**	**HD**	**WH**	**CS**
Media Geométrica	6.30	5.27	4.24	6.71	1.00	2.00	3.30	8.15	9.16	9.54	9.48	11.38
Media Armónica	6.27	5.21	4.03	6.57	1.00	2.00	3.27	7.37	9.00	9.49	9.45	11.35
Mediana	6	5	4	7	1	2	3	11	8	9	10	12
Moda	6	5	4	7	1	2	3	11	8	9	10	12
Desviación Estándar	4.67	8.67	35.00	17.67	0.00	0.00	2.67	113.67	42.67	10.92	5.00	8.92
Selección	**6º**	**5º**	**4º**	**7º**	**1º**	**2º**	**3º**	**11º**	**8º**	**9º**	**10º**	**12º**
	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×

Para el escenario 1, los resultados de las dos primeras variables, es decir, PC y PA son muy contundentes habida cuenta que, el valor numérico del *ranking* se repite para todas las pruebas tipo *single-attribute eval*uation. Las pruebas *attribute subset evaluators* consolidan dichos resultados. En cuanto al resto de las siete variables escogidas, se subraya que los valores no se repiten en todas las pruebas pero sí que se pueden interpretar fácilmente en la *tabla 7*. Todos los casos se pueden vincular a un número de ranking definitivo fácilmente. El único caso polémico es el de la variable A ya que no ha sido incluida en ninguna de las pruebas tipo *attribute subset evaluator*. Sin embargo, sus estadísticos numéricos dejan bien clara su clasificación, siendo la sexta variable más importante del primer escenario estudiado.

Escenario 2

Tabla 8: resultados la clasificación de las variables más significativas para el escenario 2 (elaboración propia).

Attribute Evaluator	Search Method	Attribute selection modes	A	E	S	C	PC	PA	PO	R	DW	HD	WH	CS
Single-Attribute Evaluators														
OneRAttributeEval	**Ranker**	Full Training set	44.17149	44.17149	44.17149	44.12822	55.00632	47.32375	44.45110	44.17149	44.17149	44.17149	44.15485	44.17149
		ranking	**7**	**5**	**4**	**12**	**1**	**2**	**3**	**8**	**10**	**9**	**11**	**6**
		Cross-validation (n=3)	44.171	44.171	44.171	44.158	54.921	47.324	44.366	44.171	44.171	44.171	44.143	44.171
		ranking	**7**	**3**	**5**	**12**	**1**	**2**	**4**	**9**	**10**	**11**	**8**	**6**
		Cross-validation (n=5)	44.171	44.171	44.171	44.166	54.921	47.324	44.423	44.171	44.171	44.171	44.153	44.171
		ranking	**9**	**4**	**5**	**7**	**1**	**2**	**3**	**10**	**12**	**8**	**11**	**6**
		Cross-validation (n=10)	44.171	44.171	44.171	44.162	54.931	47.324	44.375	44.171	44.171	44.171	44.159	44.171
		ranking	**10**	**4**	**5**	**11**	**1**	**2**	**3**	**8**	**12**	**9**	**7**	**6**
ChiSquaredAttributeEval	**Ranker**	Full Training set	723.452	1697.801	2253.960	382.664	24361.761	4503.924	2304.534	1192.513	172.078	211.764	224.491	147.096
		ranking	**7**	**5**	**4**	**8**	**1**	**2**	**3**	**6**	**11**	**10**	**9**	**12**
		Cross-validation (n=3)	491.511	1141.070	1511.495	259.087	16247.029	3011.887	1547.024	795.951	125.095	147.389	161.150	98.729
		ranking	**7**	**5**	**4**	**8**	**1**	**2**	**3**	**6**	**11**	**10**	**9**	**12**
		Cross-validation (n=5)	584.297	1364.389	1806.668	309.030	19496.921	3608.683	1846.796	955.279	142.662	173.662	186.220	118.861
		ranking	**7**	**5**	**4**	**8**	**1**	**2**	**3**	**6**	**11**	**10**	**9**	**12**

		Cross-validation (n=10)	653.865	1530.479	2030.225	345.705	21928.722	4056.304	2075.805	1073.706	157.568	192.485	205.063	132.976
		ranking	**7**	**5**	**4**	**8**	**1**	**2**	**3**	**6**	**11**	**10**	**9**	**12**
InfoGainAttributeEval	**Ranker**	Full Training set	0.01851	0.04058	0.05553	0.00986	0.42213	0.09945	0.05209	0.03061	0.00399	0.00507	0.00520	0.00392
		ranking	**7**	**5**	**3**	**8**	**1**	**2**	**4**	**6**	**11**	**10**	**9**	**12**
		Cross-validation (n=3)	0.019	0.041	0.056	0.01	0.422	0.1	0.052	0.031	0.004	0.005	0.006	0.004
		ranking	**7**	**5**	**3**	**8**	**1**	**2**	**4**	**6**	**11**	**10**	**9**	**12**
		Cross-validation (n=5)	0.019	0.041	0.056	0.01	0.422	0.1	0.052	0.031	0.004	0.005	0.005	0.004
		ranking	**7**	**5**	**3**	**8**	**1**	**2**	**4**	**6**	**11**	**10**	**9**	**12**
		Cross-validation (n=10)	0.019	0.041	0.056	0.01	0.422	0.1	0.052	0.031	0.004	0.005	0.005	0.004
		ranking	**7**	**5**	**3**	**8**	**1**	**2**	**4**	**6**	**11**	**10**	**9**	**12**
Attribute Subset Evaluators														
CfsSubseEval	**GreedyStepwise**	Full Training set	×	×	✓	×	✓	×	×	×	✓	×	×	×
		Cross-validation (n=3)	0	0	3	0	3	0	0	0	3	0	0	0
		number of folds (%)	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	5	0	5	0	0	0	3	0	0	0
		number of folds (%)	**0**	0	**100**	**0**	**100**	**0**	**0**	**0**	**60**	**0**	**0**	**0**
		Cross-validation (n=10)	0	**0**	10	0	10	0	0	0	10	0	0	0
		number of folds (%)	**0**		**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	**0**

	Exhaustive Search	Full Training set	×	×	✓	×	✓	×	×	×	✓	×	×	×
		Cross-validation (n=3)	0	0	3	0	3	0	0	0	3	0	0	0
		number of folds (%)	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	5	0	5	0	0	0	3	0	0	0
		number of folds (%)	**0**	0	**100**	**0**	**100**	**0**	**0**	**0**	**60**	**0**	**0**	**0**
		Cross-validation (n=10)	0	0	10	0	10	0	0	0	10	0	0	0
		number of folds (%)	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	**0**
	BestFirst	Full Training set	×	×	✓	×	✓	×	×	×	✓	×	×	×
		Cross-validation (n=3)	0	0	3	0	3	0	0	0	3	0	0	0
		number of folds (%)	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	**0**
		Cross-validation (n=5)	0	0	5	0	5	0	0	0	3	0	0	0
		number of folds (%)	**0**	0	**100**	**0**	**100**	**0**	**0**	**0**	**60**	**0**	**0**	**0**
		Cross-validation (n=10)	0	0	10	0	10	0	0	0	10	0	0	
		number of folds (%)	**0**	**0**	**100**	**0**	**100**	**0**	**0**	**0**	**100**	**0**	**0**	
ClassifierSubsetEval	**RandomSearch**	Full Training set	×	✓	✓	✓	×	✓	×	×	×	×	✓	✓
		Cross-validation (n=3)	0	3	3	3	0	3	0	0	0	0	3	3

	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=5)	0	5	5	5	0	5	0	0	0	0	5	5
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=10)	0	10	10	10	0	10	0	0	0	0	10	10
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**

Tabla 9: resultados de los estadísticos numéricos calculados a partir de la Tabla 8. Selección de las 7 variables más significativas para el escenario 2 (elaboración propia).

Variables	**A**	**E**	**S**	**C**	**PC**	**PA**	**PO**	**R**	**DW**	**HD**	**WH**	**CS**
Media Geométrica	7.36	4.62	3.84	8.69	1.00	2.00	3.38	6.79	10.98	9.72	9.02	9.52
Media Armónica	7.32	4.56	3.77	8.57	1.00	2.00	3.35	6.69	10.97	9.69	8.96	9.00
Mediana	7	5	4	8	1	2	3	6	11	10	9	12
Moda	7	5	4	8	1	2	3	6	11	10	9	12
Desviación Estándar	10.92	4.67	6.92	33.67	0.00	0.00	2.92	22.92	4.00	6.25	12.92	96.00
Selección	**7º**	**5º**	**4º**	**8º**	**1º**	**2º**	**3º**	**6º**	**11º**	**10º**	**9º**	**12º**
	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	×

El segundo escenario presenta resultados similares con respecto al primero, es decir, los valores *ranking* de las dos primeras variables (PC y PA) son igual de robustos para todas las pruebas. El segundo tipo de pruebas confirma también dichos resultados. En relación el resto de variables, todas ellas son también fácil de vincular un grado de importancia según los valores de los estadísticos numéricos. A pesar de que tres (A, R y PO) de las siete variables más significativas no se tuvieron en cuenta en los clasificadores de tipo *attribute subset evaluator* estos resultados se pueden entender como válidos.

Escenario 3

Tabla 10: resultados la clasificación de las variables más significativas para el escenario 3 (elaboración propia).

Attribute Evaluator	Search Method	Attribute selection modes	A	E	S	C	PC	PA	PO	R	DW	HD	WH	CS
Single-Attribute Evaluators														
OneRAttributeEval	**Ranker**	Full Training set	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	100	100	99.976	100	100	100	100	100	100	100	100	100
		ranking	**12**	**4**	**6**	**2**	**5**	**7**	**8**	**10**	**11**	**9**	**3**	**1**
		Cross-validation (n=5)	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**4**	**5**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
ChiSquaredAttributeEval	**Ranker**	Full Training set	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=5)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
InfoGainAttributeEval	**Ranker**	Full Training set	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**

		Cross-validation (n=5)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
Attribute Subset Evaluators														
CfsSubseEval	**GreedyStepwise**	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	ExhaustiveSearch	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	BestFirst	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ClassifierSubsetEval	**RandomSearch**	Full Training set	×	✓	✓	✓	×	✓	×	×	×	×	✓	✓

	Cross-validation (n=3)	0	3	3	3	0	3	0	0	0	0	3	3
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=5)	0	5	5	5	0	5	0	0	0	0	5	5
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=10)	0	10	10	10	0	10	0	0	0	0	10	10
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**

Tabla 11: resultados de los estadísticos numéricos calculados a partir de la Tabla 10. Selección de las 7 variables más significativas para el escenario 3 (elaboración propia).

Variables	**A**	**E**	**S**	**C**	**PC**	**PA**	**PO**	**R**	**DW**	**HD**	**WH**	**CS**
Media Geométrica	12.00	4.82	4.22	2.90	5.91	7.00	8.00	10.91	10.08	9.00	2.07	1.00
Media Armónica	12.00	4.80	4.19	2.88	5.90	7.00	8.00	10.91	10.08	9.00	2.06	1.00
Mediana	12	5	4	3	6	7	8	11	10	9	2	1
Moda	12	5	4	3	6	7	8	11	10	9	2	1
Desviación Estándar	0.00	1.67	4.25	0.92	0.92	0.00	0.00	0.92	0.92	0.00	0.92	0.00
Selección	**12º**	**5º**	**4º**	**3º**	**6º**	**7º**	**8º**	**11º**	**10º**	**9º**	**2º**	**1º**
	×	✓	✓	✓	✓	✓	✓	×	×	×	✓	✓

El escenario 3 presenta resultados muy robustos. La mayoría de las variables se comportan de igual modo que lo hicieron las dos primeras en los escenarios 1 y 2. Así pues, los valores numéricos del *ranking* de todas las pruebas *single-attribute evaluator* a excepción de la *OneRAttributeEval*, *Ranker*, *cross validation (n=3)*, se repiten en valor para cada variable. Las pruebas *attribute subset evaluator* confirman estos resultados a excepción de la variable PC, cuya exclusión en este tipo de pruebas es evidente. Sin embargo, no queda ningún tipo de duda que los estadísticos numéricos la clasifican como la sexta variable más importante del escenario 3.

Escenario 4

Tabla 12: resultados la clasificación de las variables más significativas para el escenario 4 (elaboración propia).

Attribute Evaluator	Search Method	Attribute selection modes	A	E	S	C	PC	PA	PO	R	DW	HD	WH	CS
Single-Attribute Evaluators														
OneRAttributeEval	**Ranker**	Full Training set	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=5)	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	100	100	100	100	100	100	100	100	100	100	100	100
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
ChiSquaredAttributeEval	**Ranker**	Full Training set	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=5)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
InfoGainAttributeEval	**Ranker**	Full Training set	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=3)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=5)	0	0	0	0	0	0	0	0	0	0	0	0

		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
		Cross-validation (n=10)	0	0	0	0	0	0	0	0	0	0	0	0
		ranking	**12**	**5**	**4**	**3**	**6**	**7**	**8**	**11**	**10**	**9**	**2**	**1**
Attribute Subset Evaluators														
CfsSubseEval	**GreedyStepwise**	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	ExhaustiveSearch	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	BestFirst	Full Training set	✓	×	×	×	×	×	×	×	×	×	×	×
		Cross-validation (n=3)	3	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=5)	5	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		Cross-validation (n=10)	10	0	0	0	0	0	0	0	0	0	0	0
		number of folds (%)	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ClassifierSubsetEval	**RandomSearch**	Full Training set	×	✓	✓	✓	×	✓	×	×	×	×	✓	✓
		Cross-validation (n=3)	0	3	3	3	0	3	0	0	0	0	3	3

	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=5)	0	5	5	5	0	5	0	0	0	0	5	5
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**
	Cross-validation (n=10)	0	10	10	10	0	10	0	0	0	0	10	10
	number of folds (%)	**0**	**100**	**100**	**100**	**0**	**100**	**0**	**0**	**0**	**0**	**100**	**100**

Tabla 13: resultados de los estadísticos numéricos calculados a partir de la Tabla 12. Selección de las 7 variables más significativas para el escenario 4 (elaboración propia).

Variables	**A**	**E**	**S**	**C**	**PC**	**PA**	**PO**	**R**	**DW**	**HD**	**WH**	**CS**
Media Geométrica	12.00	5.00	4.00	3.00	6.00	7.00	8.00	11.00	10.00	9.00	2.00	1.00
Media Armónica	12.00	5.00	4.00	3.00	6.00	7.00	8.00	11.00	10.00	9.00	2.00	1.00
Mediana	12	5	4	3	6	7	8	11	10	9	2	1
Moda	12	5	4	3	6	7	8	11	10	9	2	1
Desviación Estándar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Selección	**12º**	**5º**	**4º**	**3º**	**6º**	**7º**	**8º**	**11º**	**10º**	**9º**	**2º**	**1º**
	×	✓	✓	✓	✓	✓	×	×	×	×	✓	✓

La robustez de los resultados de las pruebas *single-attribute evaluators* de este escenario es evidente. Todas las pruebas otorgan el mismo número de *ranking* a cada variable y, en consecuencia, las desviaciones estándar son cero. Las pruebas tipo *attribute subset evaluator* apoyan estos resultados incluyendo a seis de las siete variables más importantes, incluyéndolas en al menos una de sus pruebas. De nuevo, la sexta variable (PC) no está incluida en ninguna de éstas aunque sea bien obvia su importancia dentro del escenario 4.

3.2.5. Selección de las variables predictoras más significativas

De acuerdo con los resultados obtenidos en el apartado 3.2.4, se obtienen las 7 variables predicotras en relación a la respuesta TA (*Type of Accident)*. Para ello, se ha desarrollado un total de 28 esquemas de aprendizaje en los que se ha observado: a) la posición de cada variable, b) el valor de sus estadísticos numéricos y c) la inclusión o exclusión de las variables en cada una de las pruebas de tipo *attribute subset evaluators*.

El resultado de esta metodología de trabajo ha arrojado los siguientes resultados:

Tabla 14: resultado de las 7 variables más significativas para cada uno de los 4 escenarios estudiados (elaboración propia).

Select Attributes					
Leyenda condiciones escenarios considerados			**Resultados de las 7 variables más significativas** (en orden descendente)	**nº accidentes**	**nº variable prueba**
Escenario 1	No Sobreesfuerzos	Cielo Abierto (P=3)	PC, PA, PO, S, E, A, C	8427	13
Escenario 2	No Sobreesfuerzos	Minería Subterránea (P=4)	PC, PA, PO, S, E, R, A	300042	
Escenario 3	Sobreesfuerzos	Cielo Abierto (P=3)	CS, WH, C, S, E, PC, PA	2113	
Escenario 4	Sobreesfuerzos	Minería Subterránea (P=4)	CS, WH, C, S, E, PC, PA	6658	

Como se puede observar en la *tabla 14*, los escenarios sin sobreesfuerzos (escenarios 1 y 2) toma las 7 variables más significativas prácticamente en el mismo orden de posicionamiento; la única diferencia se localiza en las dos últimas variables, cuyos resultados difieren. Para el primero, la sexta corresponde a la variable A *(Age)* y la séptima a la C (*Contract)*. En cambio, para el escenario 2 se contempla como la sexta variable R (*Risk)* mientras que la séptima A (*Age)*. En conclusión, se puede extraer que para el caso de accidentes sin sobreesfuerzo, las variables más significativas difieren escasamente. En cuanto a los escenarios que tratan los accidentes con sobreesfuerzos (escenarios 3 y 4), las 7 variables más significativas coinciden completamente tanto en presencia como en orden de importancia.

Los resultados parecen confirmar que un factor clave para determinar y comprender las causas de un accidente se obtiene mediante la discriminación entre la presencia o ausencia de un sobreesfuerzo en el suceso. A diferencia del emplazamiento, éste no parece reflejar patrones de semejanza destacables. De lo contrario, los datos registrados a cielo abierto (escenarios 1 y 4) y subterráneo (escenarios 2 y 3) producirían resultados similares, respectivamente.

3.2.6. Ejecución de árboles de decisión

Una vez seleccionadas las 7 variables más significativas, se realiza el árbol de decisión *(tree learner)* para cada uno de los escenarios propuestos. Para ello, se ha utilizado el algoritmo J4.8 cuyos resultados son los más fiables en cuanto a errores de asignación se refiere. Como variable de salida se ha tomado TA (*Type of Accident)* y el método de entrenamiento *full training set.*

En la *tabla 15* se muestran los resultados los diferentes árboles de decisión:

Tabla 15: resultado de los árboles de decisión realizados con el algoritmo J.48 (elaboración propia).

Escenario	Comparativas P=3 Vs. P=4 Sobreesfuerzo Vs. No Sobreesfuerzo	nº accidentes	nº variables predictoras	Variable salida	Evaluación prueba J48	*Size of the tree*	*Number of leaves*
Escenario 1	(1)PC (2)PA (3)PO (4)S (5)E (6)A (7)C	8427	8	TA	(56,21% / 43,79%)	245	205
Escenario 4	(1)CS (2)WH (3)C (4)S (5)E (6)PC (7)PA	6658			(100% / 0%)	1	1
Escenario 2	(1)PC (2)PA (3)PO (4)S (5)E (6)R (7)A	300042			(64,07% / 35,93%)	1007	841
Escenario 3	(1)CS (2)WH (3)C (4)S (5)E (6)PC (7)PA	2113			(100% / 0%)	1	1

Los resultados de los árboles de decisión de los escenarios sin sobreesfuerzos (escenarios 1 y 2) arrojan resultados más bien justos. Los porcentajes que representan el número de casos que se han asignado correctamente son bajos, pues para el primer escenario este valor es el 56,21% de los casos, mientras que para el segundo, es el 64,07%. En ambos casos, el tamaño del árbol aumenta en proporción con el número de accidentes que contiene cada escenario. Así pues, el tamaño del árbol de decisión del segundo escenario es más grande que el del primero.
En cuanto a los resultados de los escenarios con sobreesfuerzos (escenarios 3 y 4), llama especialmente la atención la robustez con la que se evalúan las pruebas. En este sentido, los porcentajes que representan el número de casos asignados correctamente son del 100% en ambos casos. Ello puede ser debido al tipo de codificación con el que se ha trabajado. Así pues, se evidencia que existe una fuerte relación causal entre las variables PC y TA, donde se intuye que para que tenga lugar un tipo de accidente TA=6 (sobreesfuerzo físico), previamente debe existir una causa previa de tipo PC=6 (movimiento del cuerpo con sobreesfuerzo físico). A nivel teórico, es fácil de entender que para que se produzca un accidente con sobreesfuerzo físico, previamente exista un esfuerzo físico. En otras palabras, sin esfuerzo físico no existiría sobreesfuerzo.

Como se puede apreciar, los escenarios se han presentado y ordenado en grupos de dos antagónicamente. Esto quiere decir que, por un lado se han agrupado los escenarios 1 y 4 y, por el otro, los 2 y 3. Esto permite visualizar y comparar los resultados con las condiciones opuestas de este estudio: a) ausencia o presencia

de sobresfuerzos y b) emplazamiento del accidente registrado. Se observa que para la primera agrupación, las variables que se repiten son PC (*Previous Causes)*, PA *(Physical Activity)*, S *(Size*, E *(Experience)* y A *(Age)*. Además el posicionamiento de S y E es el mismo. Para la segunda agrupación, se repiten las variables PA, S y E y, de nuevo el posicionamiento de las variables S y E coinciden.

En conclusión, se puede afirmar que las diferencias que existen entre las variables más significativas de los escenarios antagónicos demuestra la singularidad de la génesis de cada tipología de accidente.

3.2.7. Aplicación de las reglas de asociación

En minería de datos, las reglas de asociación son ampliamente practicadas para descubrir sucesos que pudieren tener lugar dentro de un determinado conjunto de datos. La confiabilidad de estas reglas determina la probabilidad con la que cada una de ellas puede ocurrir en una situación futura. Normalmente se dan por válidos aquéllos valores cuyos niveles de confiabilidad son superiores a 0.85 (85%).

Toda la información contenida en estas relaciones resulta valiosa para la toma de decisiones estratégicas en diversos ámbitos. En este estudio, se van a analizar las relaciones entre las variables más significativas que describen cada tipo de escenario. El objetivo por tanto, no es otro que el de aportar posibles soluciones que mejoren las políticas de prevención de riesgos en el sector minero español. Para ello, se ha trabajado en Weka con la regla de asociación *Predictive Apriori*, incluida dentro de la opción *Association-Rules Learners.* Se trata de un algoritmo que presenta los mejores resultados de confiabilidad ordenados en orden descendente (Sanmiquel, et al., 2015).

Como resultado de aplicar lo anteriormente expuesto, se obtiene un listado para cada escenario con sus mejores 100 reglas de asociación. Sin embargo, el presente documento únicamente recogerá las más significativas de cada escenario (ver *tablas 16,17, 18 y 19)*. Es decir, las que caracterizan y mejor explican el comportamiento de los sucesos registrados. En las tablas mencionadas, se puede apreciar como cada regla de asociación se encuentra vinculada a una condición, cuya función es la de representar un patrón que se repite en todas ellas y que, además, facilita su interpretación. Una vez expuestas, se realizará una comparativa entre los cuatro resultados obtenidos y también se realizará una evaluación de los objetivos prefijados en la introducción de este trabajo.

Por lo que se refiere al número de variables significativas utilizadas en este apartado, hay que destacar que no todas las expuestas en el apartado 3.2.5 han participado en el proceso. Ello es debido a que la inclusión de ciertas variables, repercutía en el procedimiento de cálculo arrojando resultados de compleja interpretación y cohesión. Es decir, se generaban soluciones en los que nuestra variable de trabajo respuesta predeterminada, es decir, TA (*Type of Accident)* no aparecía como una variable respuesta sino que, lo hacía en forma de causal y, además, en la mayoría de los casos. Para estudiar y explicar todos los escenarios, se ha creído conveniente realizar agrupaciones de las relaciones más interesantes en función de algunas variables. En el caso de accidentes sin sobreesfuerzo (escenarios 1 y 2) se ha hecho uso de las variables PA (*Physical Activity)* y PC *(Previous Causes)* mientras que para los casos con sobreesfuerzo (escenarios 3 y 4) la PA y WH (*Work Hours)*, ésta última como subcondición.

Escenario 1

Tabla 16: resultados de las reglas de asociación *(associate-rule learners)* más significativas para el escenario 1 (elaboración propia).

Escenario 1 (PredictiveApriori Algorithm)						
Condición	**Best Rules Found**	**Descripción**	**Confiabilidad**	**nº Atributos**	**Variables**	**nº casos**
PA=2 / PC=4	Rule 3	A=6 S=2 PC=4 PA=2 ==> TA= 2	0.98964	6	PC, PA, S, E, A, TA	8427
	Rule 12	E=4 S=2 PC=4 PA=2 ==> TA= 2	0.98487			
PC=4	Rule 7	A=3 E=3 S=1 PC=4 ==> TA=2	0.98695			
	Rule 54	A=6 E=7 S=1 PC=4 ==> TA=2	0.87114			
PA=3 / PC=4	Rule 28	E=4 S=1 PC=4 PA=3 ==> TA= 2	0.96215			
PA=5 / PC=4	Rule 34	S=3 PC=4 PA=5 ==> TA=2	0.93712			
PA=6 / PC=4	Rule 45	A=5 E=4 S=2 PC=4 PA=6 ==> TA=2	0.90421			

La *tabla 16* recoge las reglas de asociación más significativas del escenario 1. El número de variables con el que se ha trabajado es de 6 variables y el total de accidentes registrados es de 8.427. Todos ellos, sin sobreesfuerzo y a cielo abierto. Las variables PO (*Preventive Organization)* y C *(Contract)* han sido excluidas del procedimiento de cálculo por los motivos expuestos *ut supra.* La confiabilidad de todas las pruebas realizadas oscila entre el 0.87114 y 0.98964 de probabilidad, valores superiores al 0.85.

De forma particularizada se describe a continuación toda la regla de asociación de la *tabla 16:*

- *Rule 3:* empleado de 55 o más años de edad (A=6) y que trabaja en una empresa de entre 10 y 19 trabajadores (S=2) sufre un accidente en forma de golpe contra objeto estacionario (TA=2), cuando está trabajando con herramientas manuales (PA=2) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 12:* empleado con una experiencia de 61 a 120 meses (E=4) trabaja en una empresa de 10 a 19 trabajadores (S=2) y que mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2), cuando está trabajando con herramientas manuales (PA=2) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 7:* empleado de 30 a 34 años (A=3) con una experiencia de 31 a 60 meses (E=3) y que trabaja en una empresa de 0 a 9 trabajadores (S=1) mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 54:* empleado de 45 a 54 años (A=6) con una experiencia de más de 241 meses (E=7) y que trabaja en una empresa de 0 a 9 trabajadores (S=1) mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 28:* empleado con una experiencia de 31 a 60 meses (E=3) trabaja en una empresa de 0 a 9 trabajadores (S=1) mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2), cuando está conduciendo o a bordo de un medio de transporte (PA=3) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 34:* empleado que trabaja en una empresa de 20 a 49 trabajadores (S=3) mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2), cuando está transportando una carga manualmente (PA=5) y, como causa previa hubo al menos, una caída de una persona (PC=4).

- *Rule 45:* empleado de 40 a 44 años de edad (A=5) con una experiencia de 61 a 120 meses (E=4) trabaja en una empresa de 10 a 19 trabajadores (S=2) mientras sufre un accidente en forma de golpe contra un objeto estacionario (TA=2), cuando está realizando un movimiento (PA=3) y, como causa previa hubo al menos, una caída de una persona (PC=4).

De una forma generalizada se procede a extraer y describir los patrones comunes entre las diferentes reglas de asociación descritas en la *tabla 16:*

Las reglas de asociación más características de este escenario, son las que cumplen con la condición de PC=4, es decir, causa previa de accidente debido a la caída de personas. Un ejemplo de ello son las reglas 3, 12, 7 y 54, 28, 34 y 45 cuyas implicaciones siempre dan como resultado un tipo de accidente TA=2 (golpe contra objeto no móvil). Ello quiere decir que independientemente del tipo de actividad desarrollada, la causa previa debida a la caída de personas está íntimamente vinculada con la tipología de accidentes que se genera en forma de golpe contra un objeto estacionario. Además, las dos primeras también satisfacen la condición de PA=2, que implica la manipulación de objetos en el momento del accidente. Existen también otro tipo de actividades físicas que se desarrollaron mientras tenía lugar el accidente. Éstas son PA=3 (conducir o estar a bordo de un medio de transporte), PA=5 (transporte manual de cargas) y PA=6 (realización de un movimiento).

Las variable A *(Age)* y E *(Experience)* no parecen vincularse con ningún tipo de patrón fijo. En cuanto al tamaño de empresa, la variable S (*Size)* acoge una codificación variada aunque limitada a cuado S=3 (20 a 49 trabajadores). Esto quiere decir que como máximo el tamaño de la empresa no supera los 49 trabajadores. Es por tanto lógico, concluir que este tipo de relaciones tengan lugar cuando se trate de *pymes*, *p*equeñas y medias empresas.

Escenario 2

Tabla 17: resultados de las reglas de asociación *(associate-rule learners)* más significativas para el escenario 2 (elaboración propia).

Escenario 2 (PredictiveApriori Algorithm)						
Condición	**Best Rules Found**	**Descripción**	**Confiabilidad**	**nº Atributos**	**Variables**	**nº casos**
PC=2	Rule 23	A=4 E=4 S=3 PC=2 ==> TA=3	0.99302	6	PC, PA, S, E, A, TA	30042
	Rule 94	A=3 E=1 S=4 PC=2 ==> TA=3	0.95191			
PA=2 / PC=2	Rule 3	A=4 E=1 S=4 PC=2 PA=2 ==> TA=3	0.99451			
	Rule 60	A=1 E=2 S=5 PC=2 PA=2 ==> TA=3	0.9861			
PA=4 / PC=2	Rule 78	A=3 E=5 S=5 PC=2 PA=4 ==> TA=3	0.97759			
	Rule 54	A=2 E=4 S=6 PC=2 PA=4 ==> TA=3	0.98712			

La *tabla 17* recoge las reglas de asociación más significativas del escenario 2. La cantidad y uso de variables utilizadas es la misma que para el escenario 1, 6 variables aunque con un mayor número de accidentes registrados que asciende hasta los 30.042. Las variables PO (*Preventive Organization)* y R *(Risk)* han sido excluidas del procedimiento de cálculo por los motivos expuestos *ut supra.* En esta ocasión, ambas variables interferían en las implicaciones de todas las reglas de asociación, excluyendo a TA de la respuesta, es decir, de su aparición en la parte derecha de la flecha. La confiabilidad de todas las pruebas realizadas oscila entre el 0.95191 y 0.99451 de probabilidad, valores altos y superiores al 0.85.

De forma particularizada se describe a continuación toda la regla de asociación de la *tabla 17:*

- *Rule 23:* empleado de 35 y 39 años de edad (A=4) y una experiencia de 50 a 99 meses (E=3) que trabaja en una empresa de entre 20 y 49 trabajadores (S=3) sufre un accidente en forma de golpe contra objeto móvil (TA=3), cuando como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

- *Rule 94:* empleado de entre 30 y 34 años de edad (A=3) y una experiencia de 0 a 9 meses (E=1) trabaja en una empresa de 50 a 99 trabajadores (S=4) mientras sufre un accidente en forma de golpe contra un objeto móvil (TA=3), cuando como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

- *Rule 3:* empleado de entre 16 y 24 años de edad (A=1) con una experiencia de 0 a 9 meses (E=1) trabaja en una empresa de 50 a 99 trabajadores (S=4) mientras sufre un accidente en forma de golpe contra un objeto móvil (TA=3), cuando está trabajando con herramientas manuales (PA=2) y, como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

- *Rule 60:* empleado de entre 35 y 39 años de edad (A=4) con una experiencia de 10 a 19 meses (E=2) trabaja en una empresa de 100 a 499 trabajadores (S=5) mientras sufre un accidente en forma de golpe contra un objeto móvil (TA=3), cuando está trabajando con herramientas manuales (PA=2) y, como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

- *Rule 78:* empleado de entre 30 y 34 años de edad (A=3) con una experiencia de 100 a 499 meses (E=5) trabaja en una empresa de 100 a 499 trabajadores (S=5) mientras sufre un accidente en forma de golpe contra un objeto móvil (TA=3), cuando está manipulando objetos (PA=4) y, como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

- *Rule 54:* empleado de entre 25 y 29 años de edad (A=2) con una experiencia de 61 a 120 meses (E=4) trabaja en una empresa de más de 500 trabajadores (S=6) mientras sufre un accidente en forma de golpe contra un objeto móvil (TA=3), cuando está manipulando objetos (PA=4) y, como causa previa hubo una rotura, un estallido, una fuga o derrumbamiento de materiales (PC=2).

De una forma generalizada se procede a extraer y describir los patrones comunes entre las diferentes reglas de asociación descritas en la *tabla 17:*

Todas las reglas de asociación encontradas cumplen de forma general con la condición PC=2 (causa previa de rotura, estallido, fuga o derrumbamiento de agente material) y, en algunos casos también lo hacen cuando de forma simultánea cumplen las de PA=2 (trabajos con herramientas manuales) y PA=4 (manipulación de objetos), respectivamente. Las implicaciones de estas tres condiciones son las misma, TA=3 (accidente debido a un golpe contra un objeto móvil). Se deduce por tanto, una fuerte relación entre la causa previa tipo PC=2 y el tipo de accidente TA=3.

En cuanto al resto de variables, los valores de A (*Age)* y E *(Experience)* son de nuevo variados y parecen seguir un patrón estable determinado. Sin embargo, sí que lo sigue la variable S *(Size)*, cuyos valores suele ser más elevados que los del escenario 1. Así pues, se evidencia que este tipo de relaciones son frecuentes cuando se trata de empresas con un número de trabajadores de entre 20 y 49 (S=3), 50 y 99 (S=4), 100 y 499 (S=5) y S=6 (500 o más), siendo más frecuentes las tres últimas franjas, es decir, empresas de un mayor tamaño.

Escenario 3

Tabla 18: resultados de las reglas de asociación *(associate-rule learners)* más significativas para el escenario 3 (elaboración propia).

Escenario 3 (PredictiveApriori Algorithm)						
Condición	**Best Rules Found**	**Descripción**	**Confiabilidad**	**nº Atributos**	**Variables**	**nº casos**
PA=4	Rule 14	PA=4 WH=3 ==> PC=6 TA=5	0.99499	6	WH, S, E, PC, PA, TA	2113
	Rule 22	S=1 PA=4 ==> PC=6 TA=5	0.99498			
PA=2	Rule 34	PA=2 WH=3 ==> PC=6 TA=5	0.99496			
	Rule 39	S=3 PA=2 ==> PC=6 TA=5	0.99495			
PA=5	Rule 44	PA=5 WH=3 ==> PC=6 TA=5	0.99494			
	Rule 50	S=2 PA=5 ==> PC=6 TA=5	0.99493			
PA=6	Rule 20	PA=6 WH=3 ==> PC=6 TA=5	0.99498			
	Rule 46	S=1 PA=6 ==> PC=6 TA=5	0.99494			

La *tabla 18* recoge las reglas de asociación más significativas del escenario 3. El número de variables con el que se ha trabajado es superior al de los dos escenarios previos, 7 variables un total de 2.113 accidentes registrados. Las variables C *(Contract)* y CS *(Contractual Status)* se han excluido del procedimiento de cálculo por los motivos ya expuestos. La confiabilidad de todas las pruebas realizadas oscila entre el 0.99087 y 0.995 de probabilidad, un rango ligeramente superior al de los dos escenarios previos.

De forma particularizada se describe a continuación toda la regla de asociación de la *tabla 18:*

- *Rule 14:* un empleado que lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está manipulando objetos (PA=4) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 22:* un empleado que trabaja en una empresa de 0 a 9 trabajadores (S=1) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está manipulando objetos (PA=4) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 34:* un empleado que lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está trabajando con herramientas manuales (PA=2) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 39:* un empleado que trabaja en una empresa de 20 a 49 trabajadores (S=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está trabajando con herramientas manuales (PA=2) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 44:* un empleado que lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está transportando una carga manualmente (PA=5) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 50:* un empleado que trabaja en una empresa de 10 a 19 trabajadores (S=2) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está transportando una carga manualmente (PA=5) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 20:* un empleado que lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está realizando un movimiento (PA=6) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 46:* un empleado que trabaja en una empresa de 0 a 9 trabajadores (S=1) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está realizado un movimiento (PA=6) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

De una forma generalizada se procede a extraer y describir los patrones comunes entre las diferentes reglas de asociación descritas en la *tabla 18:*

Las reglas de asociación más significativas de este caso son la 14 y 22 cuando cumplen la condición de PA=4 (manipulación de objetos), la 34 y 56 cuando PA=2 (trabajos con herramientas manuales), la 44 y 50 cuando PA=5 (transporte manual de cargas) y la 20 y 46 cuando PA=6 (realización de un movimiento). Todas ellas a dan como variables respuesta un accidente de tipo TA=5 (sobreesfuerzo físico) y una causa previa PC=6 (movimiento del cuerpo con esfuerzo físico).

Las variables E *(Experience)* es irrelevante para este escenario habida cuenta su ausencia en todas y cada una de las reglas propuestas en la *tabla 18*. Sin embargo, las variables S *(Size)* y WH *(Work Hours)* sí que toman el protagonismo en aquéllas cuando se observa que el tamaño de empresa predilecto es más bien pequeño, es decir, cuando S=1, S=2 y S=3. En cuanto al número de horas trabajadas se refiere, en la mitad de las reglas aparece WH=3, lo que significa el empleado sufrió el accidente cuando acumulaba un total de entre 4 y 8 horas de faena completadas.

Escenario 4

Tabla 19: resultados de las reglas de asociación *(associate-rule learners)* más significativas para el escenario 4 (elaboración propia).

Escenario 4 (PredictiveApriori Algorithm)						
Condición	**Best Rules Found**	**Descripción**	**Confiabilidad**	**nº Atributos**	**Variables**	**nº casos**
PA=4	Rule 89	S=5 PC=6 PA=4 WH=1 ==> TA=5	0.99496	6	WH, S, E, PC, PA, TA	6658
	Rule 85	E=4 S=5 PA=4 WH=3 =0 ==> PC=6 TA=5	0.99499			
PA=2	Rule 5	S=5 PA=2 WH=2 =0 ==> PC=6 TA=5	0.99499			
	Rule 40	E=4 S=6 PA=2 ==> PC=6 TA=5	0.99498			
PA=5	Rule 31	S=5 PA=5 WH=3 ==> PC=6 TA=5	0.99499			
	Rule 6	E=4 PA=5 =0 ==> PC=6 TA=5	0.99499			
PA=6	Rule 12	PA=6 WH=2 ==> PC=6 TA=5	0.99499			

La *tabla 19* recoge las reglas de asociación más significativas del escenario 4. Tanto la cantidad como el uso de variables utilizadas es la misma que para el escenario 3, 7 variables con un mayor número de datos, 6.658 accidentes registrados. Las variables C *(Contract)* y CS *(Contractual Status)* también se ha excluido del procedimiento. La confiabilidad de todas las pruebas realizadas oscila entre 0.99496 y 0.99499, superior al 0.85 establecido.

De forma particularizada se describe a continuación toda la regla de asociación de la *tabla 19:*

- *Rule 89:* un empleado que trabaja en una empresa de 100 a 499 trabajadores (S=5) y lleva una jornada laboral completada de entre 0 y 1 hora (WH=1) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está manipulando objetos (PA=4) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 85:* un empleado con una experiencia de 61 a 120 meses (E=4) que trabaja en una empresa de 100 a 499 trabajadores (S=5) y lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está manipulando objetos (PA=4) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 5:* un empleado que trabaja en una empresa de 100 a 499 trabajadores (S=5) y lleva una jornada laboral completada de entre 1 y 4 horas (WH=2) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está trabajando con herramientas manuales (PA=2) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 40:* un empleado con una experiencia de 61 a 120 meses (E=4) que trabaja en una empresa de más de 500 trabajadores (S=6) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está trabajando con herramientas manuales (PA=2) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 31:* un empleado trabaja en una empresa de 100 a 499 trabajadores (S=5) y lleva una jornada laboral completada de entre 4 y 8 horas (WH=3) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está transportando una carga manualmente (PA=5) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 6:* un empleado con una experiencia de 61 a 120 meses (E=4) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está transportando una carga manualmente (PA=5) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

- *Rule 12:* un empleado que lleva una jornada laboral completada de entre 1 y 4 horas (WH=2) sufre un accidente en forma de sobreesfuerzo físico (TA=5) cuando está realizando un movimiento (PA=6) y, como causa previa éste realizó un movimiento del cuerpo con esfuerzo físico (PC=6).

De una forma generalizada se procede a extraer y describir los patrones comunes entre las diferentes reglas de asociación descritas en la *tabla 19:*

Las reglas de asociación singulares de esta casuística, son la 89 y 85 cuando PA=4 (manipulación de objetos), la 5 y 40 cuando PA=2 (trabajos con herramientas manuales), la 31 y 6 cuando PA=5 (transporta manual de cargas) y la 12 y 27 cuando PA=6 (realización de un movimiento). Las implicaciones de las mismas dan como resultado un tipo de accidente TA=5 (sobreesfuerzo físicos) y una causa previa PC=6 (movimiento con esfuerzo físico).

En cuanto al resto de variables se refiere, la variable E *(Experience)* aparece en varios de los casos expuestos cuando E=4, experiencia del trabajador comprendida entre los 61 y 120 meses. El caso de la variable S *(Size)* WH *(Work Hours)* sigue los patrones establecidos en el escenario 2. Por un lado, empresas grandes con 100 o más trabajadores (S=5 y S=6) y, por el otro, jornadas laborables completadas de no más de 8 horas (WH=1, WH=2 y WH=3).

3.2.8. Comparativa de los resultados obtenidos

Una vez obtenidas las reglas de asociación más significativas de cada escenario, conviene destacar las semejanzas y diferencias entre los patrones de comportamiento obtenidos.

Así pues se procede a comparar en bloques de dos, cada uno de los escenarios estudiados:

Escenario 1 Vs. Escenario 2

Estos dos escenarios se caracterizan por registrar accidentes en los que no hubo sobreesfuerzo. Sin embargo, el primero hace referencia a aquéllos que ocurrieron a cielo abierto mientras que, el segundo a los que sucedieron en minería subterránea.

La variable TA es diferente para ambos casos, es decir, el tipo de accidente varía en función del emplazamiento. Así pues, a cielo abierto se observa que TA=2 (golpe contra objeto estacionario) mientras que para minería subterránea TA=3 (golpe contra objeto móvil).

La variable PC parece ser una de las que se encuentra íntimamente ligadas con el tipo de accidente. En consecuencia, para el escenario 1, la causa previa es igual PC=4 (caídas de personas) y para el escenario 2, la PC=2 (rotura, estallido, fuga o derrumbamiento de agente material). De acuerdo con las causas previas, el tipo de accidentes mencionados en el párrafo anterior, todavía adquieren mayor sentido, es decir, una caída de una persona normalmente se traducirá en un golpe contra un objeto que bien pudiera estar a su alrededor y que además no estuviera en movimiento. Lo mismo sucede con el escenario 2, pues si tiene lugar una fuga, un estallido o un derrumbamiento del agente material, es lógico que el tipo de accidente sea en forma de golpe contra un objeto en movimiento.

En cuanto al tamaño de empresa (variable S) se observan diferentes resultados para ambos escenarios, siendo más habitual encontrar empresas pequeñas en el escenario 1 (S=1,2 o 3) y empresas de mayor tamaño en el escenario 2 (S=3,4, 5 o 6). Estos resultados se corresponden con la realidad habida cuenta que, la mayoría de empresas que se dedican a la minería subterránea suelen ser siempre mayores que las que se dedican a la minería a cielo abierto.

El valor de la variable PA, actividad física desarrollada es algo más variado. Para el primer escenario esta es igual a PA=2 (trabajos con herramientas manuales), PA=3 (conducir o estar a bordo de un medio de transporte), PA=5 (transporte manual de cargas) o PA=6 (realización de un movimiento). En el caso de la minería subterránea, dicha variable es igual a PA=2 o PA=4 (manipulación de objetos). Parece ser que todas ellas se encuentran relacionadas con los tipos de accidente TA=2 y TA=3. Además, se observa que el rango de actividades desarrolladas a cielo abierto es mayor que en el de la minería subterránea.

En cuanto a las variables E y A no se ha encontrado ningún patrón diferente o en común. Los valores son inestables y por tanto, no se puede extraer ninguna conclusión al respecto.

Escenario 3 Vs. Escenario 4

Estos dos escenarios se caracterizan por registrar accidentes en los que hubo un sobreesfuerzo. Sin embargo, el escenario 3 hace referencia a aquéllos que ocurrieron a cielo abierto mientras que, el escenario 2 engloba a los que sucedieron en minería subterránea.

La variable TA y PC es la misma para ambos casos, es decir, TA=5 (sobreesfuerzo físico) y PC=6 (movimiento del cuerpo con esfuerzo físico). Esta relación es totalmente lógica ya que para que exista un sobreesfuerzo previamente debe haber la realización de un esfuerzo físico. Las reglas de asociación de las tablas *18 y 19* determinan que existe una fuerte relación entre PC y TA.

Con respecto a la variable S, se observa que el patrón que caracteriza el tamaño de las empresas varía en función del tipo de minería al que se dedican. Por tanto, las empresas de minería a cielo abierto son pequeñas y las de minería subterránea más grandes, independientemente de si hubo o no un sobreesfuerzo en el accidente.

Las codificaciones del tipo de actividad física desarrollada para el escenario 3 y 4 son las mismas. Se observa por tanto, que PA=2, P=4, PA=5 y PA=6 en ambos casos. Esta comparativa puede interpretarse de manera que cuando TA=6 (tipo de accidente de sobreesfuerzo) y necesariamente PC=6, entonces el rango de actividades desarrolladas tanto a cielo abierto como en minería subterránea es el mismo. En otras palabras, esto quiere decir que si se registra un accidente con sobreesfuerzo, éste independientemente del emplazamiento, depende del tipo de actividad física

desarrollada. Es decir, que con sobreesfuerzos la tipología de accidentes a cielo abierto o subterráneo es muy parecida.

El caso de la variable WH es fácil de interpretar ya que para ambos escenarios el valor que ésta adquiere es 1, 2 o 3. Esto quiere decir que los accidentes suelen ocurrir a primera hora de la mañana. Conviene recordar que los mineros empiezan su jornada laboral bien temprano y que, después de almorzar o comer (WH=2 o WH=3) es cuando más probabilidad hay de que tenga lugar un accidente. Estas franjas horarias representan por tanto, los momentos en los que todavía el empleado puede estar realizando parte de la digestión o más relajado y desactivado se encuentra.

Escenario 1 Vs. Escenario 4

Estos dos escenarios se caracterizan por registrar accidentes completamente opuestos según: a) el tipo de emplazamiento y, b) la presencia o ausencia de sobreesfuerzos. En consecuencia, las 7 variables más significativas de cada escenario no coinciden.

De entre las variables comunes en ambos escenarios encontramos:

La variable TA es común en ambos casos debido a que se trata de la variable respuesta. Se observan tipologías de accidentes bien diferentes en ambos casos. En el escenario 1, TA=2 y en el 4, TA=5.

La variable PC es también común aunque asuma valores diferentes para cada escenario. En el escenario 1 ésta siempre es PC=4 (caídas de personas) mientras que para el escenario 4 es PC=6 (movimiento con esfuerzo físico). En consecuencia, se aprecia que la génesis de los accidentes en relación a la causa previa es bien diferente según el emplazamiento y la valoración del sobreesfuerzo del accidente.

El tipo de actividad física desarrollada (PA) sí que comparte algunas codificaciones entre escenarios opuestos. En concreto, las que se refieren a cuando PA=2 (trabajos con herramientas manuales), PA=5 (transporte manual de cargas) y, PA=6 (realización de un movimiento). Todas ellas fácilmente reproducibles en cualquiera de los dos escenarios. Sin embargo, el tipo de implicación que ello conlleva -como ya se ha explicado- no se traduce en los mismos resultados, pues así lo refleja la variable TA.
La variable S parece seguir los patrones ya explicados en las dos comparativas previas. Es decir en los escenarios donde se contemplan accidentes a cielo abierto, el tamaño de empresa es más pequeño mientras que, cuando se aprecian los de minería subterránea, las empresas incluyen un mayor número de trabajadores. Así también lo constatan los resultados: escenario 1 (S=1,2 y 3) y escenario 4 (S=5 y 6).

En cuanto a la variable experiencia (E), los resultados arrojan codificaciones múltiples y, por tanto, se trata de un factor más complejo de interpretar, al menos, desde el enfoque del presente estudio.

De entre las variables no comunes entre los escenarios 1 y 4 se encuentran WH y A.

Escenario 2 Vs. Escenario 3

Estos dos escenarios se caracterizan por registrar accidentes completamente opuestos según: a) el tipo de emplazamiento y, b) la presencia o ausencia de sobreesfuerzos. En consecuencia, las 7 variables más significativas de cada escenario no coinciden.

De entre las variables comunes en ambos escenarios encontramos:

La variable TA es común en ambos casos debido a que se trata de la variable respuesta. Se observan tipologías de accidentes bien diferentes en ambos casos. En el escenario 2, TA=3 y en el 3, TA=5.

La variable PC es también común aunque asuma valores diferentes para cada escenario. En el escenario 2 ésta siempre es PC=2 (rotura, estallido, fuga o derrumbamiento de agente material) mientras que para el escenario 3 es PC=6 (movimiento con esfuerzo físico). En consecuencia, se aprecia –de igual modo que en la comparativa anterior- que la génesis de los accidentes en relación a la causa previa es bien diferente según el emplazamiento y la valoración del sobreesfuerzo del accidente.

El tipo de actividad física desarrollada (PA) sí que comparte algunas codificaciones entre escenarios opuestos. En concreto, las que se refieren a cuando PA=2 (trabajos con herramientas manuales), PA=4 (manipulación de objetos).

La variable S parece sigue los patrones de las comparativas previas y así lo corroboran los resultados obtenidos: escenario 2 (S=3, 4, 5 y 6) y escenario 3 (S=1, 2 y 3).

En cuanto a la variable experiencia (E), los resultados arrojan codificaciones múltiples y, por tanto, se trata de un factor más complejo de interpretar, al menos, desde el enfoque del presente estudio.

De entre las variables no comunes entre los escenarios 2 y 3 se encuentran la WH y A, las mismas que para el caso anterior.

3.3. Evaluación de los resultados

Los resultados obtenidos parecen confirmar que:

1. El tipo de accidente (TA) más habitual en los escenarios sin sobreesfuerzos son debidos a golpes contra objetos estacionarios o móviles y que para escenarios con sobreesfuerzos, aquéllos son sobreesfuerzos físicos sobre el sistema musculo esquelético.

2. Las causas previas del accidente (PC) es un factor que se encuentra íntimamente ligado a la tipología de accidentes acontecido. En este sentido, una causa previa de clase 4, 2 y 6 siempre da lugar a un tipo de accidente 2, 3 y 5, respectivamente.

3. La actividad física desarrollada (PA) es una variable común en todos los escenarios y que, además acoge los mismos valores en algunos de los casos. Es por tanto, una variable que depende fuertemente de la causa previa al accidente.

4. El tamaño de la empresa (S) que se ve implicado en estas reglas de asociación depende de la tipología de minería al que se dedique la misma. Para cielo abierto empresas pequeñas, para minería subterránea empresas grandes.

5. Las horas de jornada laboral (WH) en las que habitualmente se producen los accidentes son aquéllas de tipo 1, 2 y 3, correspondiendo con las horas después del descanso, desayuno o comida.

6. Las variables referentes a la edad (A) y la experiencia (E) del trabajador no arrojan resultados estables que expliquen una cierta casuística concreta.

7. Las variables de evaluación de riesgo (R), organización preventiva (PO) y tipología de contratación (CS) producían problemas en la interpretación y la cohesión de las reglas de asociación resultantes.

8. Las variables más significativas de los escenarios 1 y 2 (sin sobreesfuerzos) son A, E, S, PC, PA y TA.

9. Las variables más significativas de los escenarios 3 y 4 (con sobreesfuerzos) son E, S, PC, PA, WH y TA.

4. Estadística descriptiva

4.1. Introducción

La estadística descriptiva o análisis exploratorio de datos tiene como objetivo ofrecer modos de presentar y evaluar las características principales de los datos a través de tablas de frecuencia, gráficos de barras y comentarios.

El objetivo principal de los gráficos es el de representar los datos de manera que se puedan apreciar como un todo e identificar sus características sobresalientes, sobretodo, cuando se trata con una potente base de datos. El tipo de gráfico a seleccionar varía según el tipo de variable que interese evaluar. Es por tanto lógico que, previamente al análisis descriptivo se proceda a identificar el tipo de datos con el que se está trabajando. Ello determina el método de análisis más apropiado y válido para el conjunto de datos. Así pues, conviene subrayar que una de las distinciones más importantes es la que se da entre datos numéricos y categóricos.

Las variables de nuestra base de datos son categóricas o cualitativas. Ello quiere decir que cada número registra la presencia de un atributo. En este sentido, es importante destacar que las categorías de una variable cualitativa deben ser definidas claramente durante la etapa de diseño de datos de la investigación y que, en consecuencia, éstas acaben siendo exhaustivas y mutuamente excluyentes. Por consiguiente, cada unidad de observación debe clasificarse sin ambigüedad en una y sólo una de las categorías posibles siendo posible a su vez, clasificar a todo individuo sin excepción alguna.

Los datos categóricos o cualitativos se clasifican en dos categorías: a) dicotómicos, b) politómicos.

a) **Dicotómicos**: son datos que se registran en dos condiciones. El individuo o la unidad de observación puede ser asignada a solo una de dos categorías. En general se trata de la dicotomía entre *presencia* – *ausencia* del atributo y suele resultar ventajoso asignar el código 0 a la ausencia y el 1 a la presencia.

b) **Politómicos:** son datos que se registran en tres o más condiciones, es decir, cuando la unidad de observación puede adquirir tres o más valores.

b.1) Ordinales: son datos que se almacenan siguiendo un orden natural entre las categorías. Normalmente suelen ser escalas establecidas aunque el intervalo de medición no tiene por qué ser necesariamente uniforme.

b.2) Nominales: son datos que no pueden ser sometidos a un criterio de orden, es decir, no existe un orden obvio entre las categorías.

Se debe ser cuidadoso cuando se trabaja con variables cualitativas que se han codificado numéricamente puesto que no pueden ser analizadas como números, sino que deben ser analizadas como categorías.

El fichero con el que se trabaja en *Minitab v.16* cuenta con un total de 13 variables categóricas de las cuales doce son predictoras y únicamente una es respuesta. Así que, de acuerdo con lo expuesto anteriormente, la manera más simple de presentar una base de datos con variables categóricas es mediante un gráfico de barras. Esta agrupación de datos se realiza mediante el uso de una tabla de frecuencias en la que se indica el número de unidades de análisis que caen en cada una de las clases de la variable cualitativa. La representación gráfica de una distribución de frecuencias puede realizarse a través de un histograma, gráfico de barras o un gráfico de tortas.

4.2. Descripción de los datos

Como ya se ha mencionado en el apartado 3.2.2 este apartado se centra en el estudio de la fiabilidad de los datos. Para ello se dispondrá de las herramientas de la estadística descriptiva. Las variables a estudiar en este apartado, serán aquellas que se han considerado como significativas en las reglas de asociación propuestas en las *tablas 16, 17, 18 y 19.* En cuanto al resto de variables, cabe destacar que quedan al margen de este estudio descriptivo.

4.2.1. Accidentes registrados sin sobreesfuerzos

Los accidentes sin sobreesfuerzo son contemplados en los escenarios 1 y 2 del presente trabajo y, las variables a estudiar serán A, E, S, PC, PA y TA.

1) A *(Age)*

Tabla 20: Tabla comparativa de frecuencias de la variable edad (A) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

A	Años	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	16-24	796	796	9,4458	9,44	841	841	2,7994	2,799
2	25-29	978	1774	11,6056	21,051	2692	3533	8,9608	11,760
3	30-34	1285	3059	15,2486	36,300	4770	8303	15,8778	27,638
4	35-39	1323	4382	15,6995	52,000	9348	17651	31,1164	58,754
5	40-44	1311	5693	15,5571	67,557	10119	27770	33,6828	92,437
6	45-54	1892	7585	22,4516	90,008	2162	29932	7,1966	99,634
7	55 o más	842	8427	9,9917	100,000	110	30042	0,3662	100,000

– El número de casos registrados (N) de la *tabla 20* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

– La **centralidad** de los datos para ambos emplazamientos se encuentra comprendida entre las clases 3 y 4, es decir, empleados entre los 30 y 34 años de edad. Contrariamente, se observa que la clase modal para cada estadio no es la misma. Para el caso de accidentes mineros

producidos a cielo abierto, la clase más observada es la 6, trabajadores de entre 45 y 54 años mientras que, para los casos de minería subterránea aquélla pasa a ser la 5, trabajadores de 40 a 44 años. De esta forma, se traducen también los porcentajes representados en la *tabla 20*, donde a cielo abierto la clase 5 acumula el 33,68% de los accidentes y en el caso de la minería subterránea la clase 6 significa el 22,45% de los datos registrados.

– En cuanto a la **distribución** de los datos, se observa que para ambos emplazamientos ésta sigue una asimetría negativa. En el primer estadio, se observa que la forma de la distribución de los datos es de picos bajos, pues las clases 3, 4, 5 y 6 son los picos de la distribución y, sin embargo, asumen valores parecidos y que se alejan muy poco del resto de las clases. En el segundo estadio, el grado de apilamiento de los datos alrededor de un punto se asemeja más a la de una distribución normal. Los picos de esta segunda distribución, clases 4 y 5, sí se diferencian claramente del resto.

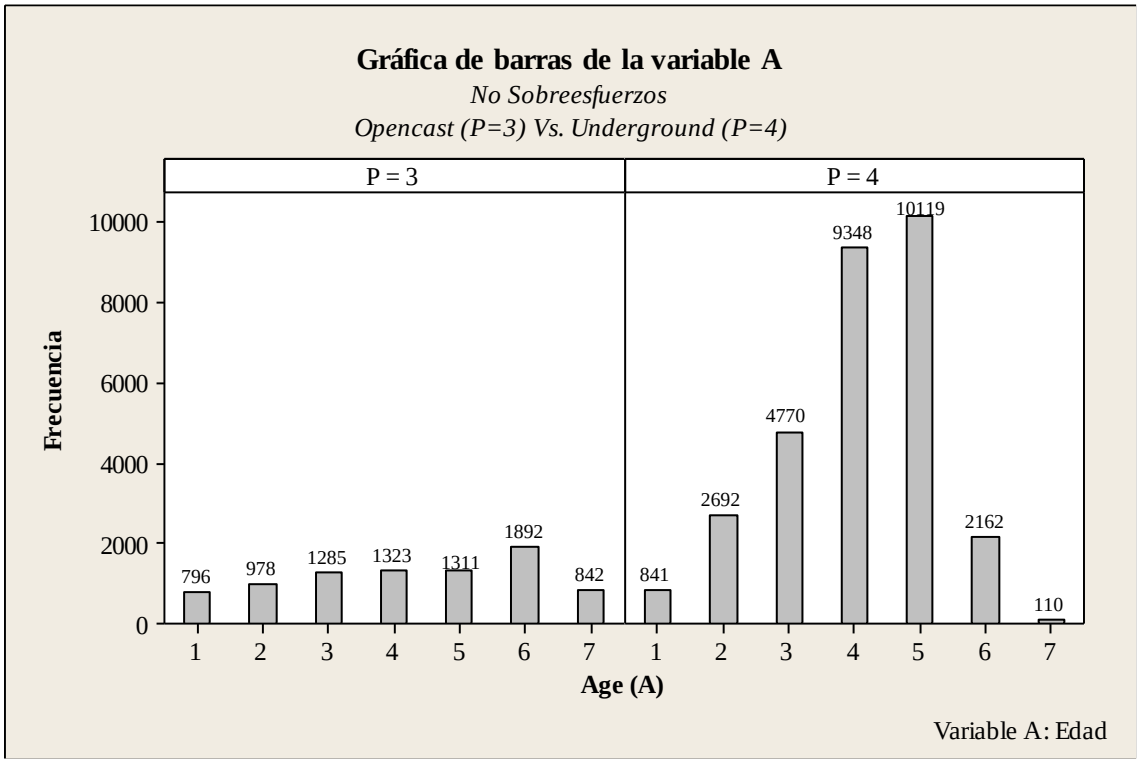

Figura 2: Gráfico de barras comparativo de la variable edad (A) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

2) E *(Experience)*

Tabla 21: Tabla comparativa de frecuencias de la variable experiencia (E) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

E	Meses	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	0-12	3010	3010	35,7185	35,719	4831	4831	16,0808	16,081
2	13-30	1470	4480	17,4439	53,162	4484	9315	14,9258	31,007
3	31-60	1340	5820	15,9013	69,064	3579	12894	11,9133	42,920
4	61-120	1599	7419	18,9747	88,038	9491	22385	31,5924	74,512
5	121-180	469	7888	5,5654	93,604	2705	25090	9,0041	83,516
6	181-240	214	8102	2,5395	96,143	3634	28724	12,0964	95,613
7	241, más	325	8427	3,8567	100,000	1318	30042	4,3872	100,000

– El número de casos registrados (N) de la *tabla 21* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

– La **centralidad** de los datos para ambos emplazamientos es diferente. Para cielo abierto ésta se encuentra comprendida entre las clases 1 y 2 mientras que para minería subterránea se localiza entre las clases 3 y 4. El tercer cuartil del primer caso corresponde a la clase 4 y evidencia el poco peso que aportan las tres últimas clases del gráfico de barras. Para el segundo caso, el tercer cuartil coincide con la categoría 5. La clase modal del primer escenario es la 1 mientras que para la el segundo es la 5.

– Acerca de la **distribución** de los datos, se observan asimetrías positivas en ambos emplazamientos a pesar de que ésta se encuentre mucho más acentuada en el primer escenario que no en el segundo. Los picos de la primera clase son más bien bajos, destacando el de la clase 1; de entre los picos de las segunda, destaca el de tipo 4, la clase modal del escenario 2.

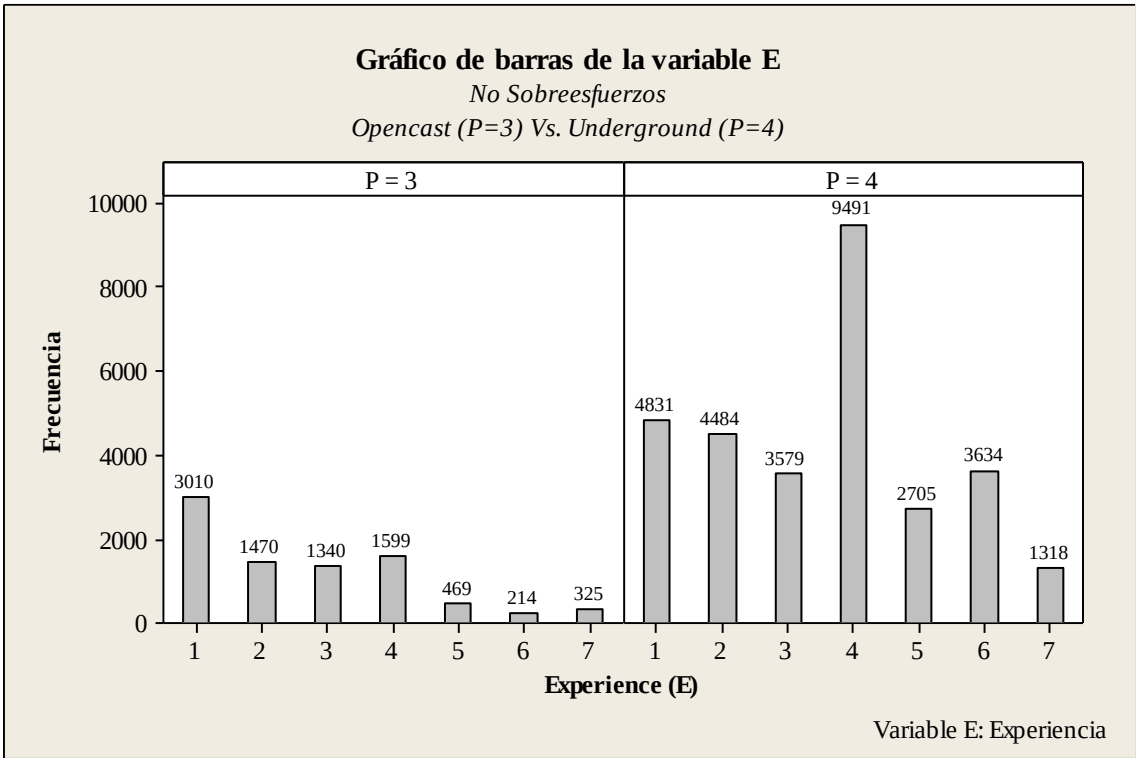

Figura 3: Gráfico de barras comparativo de la variable experiencia (E) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

3) S *(Size)*

Tabla 22: Tabla comparativa de frecuencias de la variable tamaño de empresa (S) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

S	Número de empleados	**Lugar accidente** *(place)*							
		Cielo abierto *(3_opencast)*				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	0-9	2340	2340	27,7679	27,768	508	508	1,6910	1,691
2	10-19	2257	4597	26,7830	54,551	549	1057	1,8274	3,518
3	20-49	2490	7089	29,5716	84,122	2720	3777	9,0540	12,572
4	50-99	682	7771	8,0930	92,215	2720	6497	9,0540	21,626
5	100-499	624	8395	7,4048	99,620	14863	21360	49,4741	71,100
6	500 o más	32	8427	0,3797	100,000	8682	30042	28,8995	100,000

- El número de casos registrados (N) de la *tabla 22* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

- La **centralidad** de los datos es diferente para el tipo de emplazamiento analizado. Para el caso de cielo abierto, está se encuentra contenida entre las clases 1 y 2 mientras que para la minería subterránea, ésta se localiza entre las categorías 4 y 5. La clase modal del escenario 1 es la 3 y la del escenario 2 es la 5, representando el 29,57% y 49,47% de los datos, respectivamente. El tercer cuartil a cielo abierto se ubica en la clase 3 mientras que en minería subterránea corresponde a la clase 5.

– En cuanto a la **distribución** de los datos de la variable S, se observan comportamientos opuestos entre P=3 y P=4. No se aprecia ningún tipo de simetría y por tanto, la primera distribución corresponde a una asimetría positiva y la segunda a una negativa. Los picos del gráfico de barras cuando P=3 son la clase 1, 2 y 3. En cambio, cuando P=4, los picos se localizan en las clases 5 y 6.

– Los resultados se corresponden con las hipótesis efectuadas en el apartado 3.3.

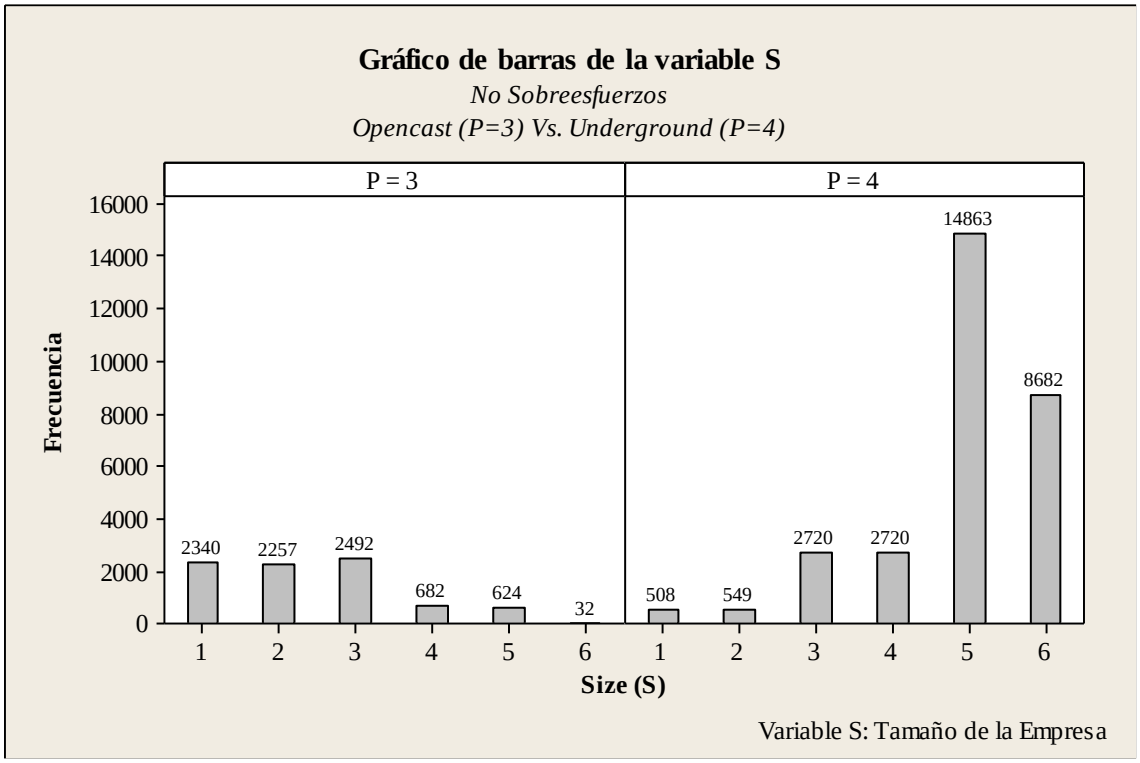

Figura 4: Gráfico de barras comparativo de la variable tamaño de empresa (S) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

5) PC *(Previous Causes)*

Tabla 23: Tabla comparativa de frecuencias de la variable causas previas (PC) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

PC	Tipo de causa previa	**Lugar accidente *(place)***							
		Cielo abierto (*3_opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	Problema eléctrico, explosión, fuego, desbordamiento, vuelco, escape, derrame, vaporización	452	452	5,3637	5,364	1137	1137	3,7847	3,785
2	Rotura, fractura, estallido, resbalón, caída, derrumbamiento de agente material	1277	1729	15,1537	20,517	11250	12387	37,4476	41,232
3	Pérdida de control (total o parcial) de control de máquinas	1725	3454	20,4699	40,987	7905	20292	26,3132	67,545
4	Caídas de personas	1898	5352	22,5228	63,510	4132	24424	13,7541	81,300
5	Movimiento del cuerpo sin esfuerzo físico	1855	7207	22,0126	85,523	2791	27215	9,2903	90,590
6	Movimiento del cuerpo con esfuerzo físico	607	7814	7,203	92,726	1337	28552	4,4504	95,040
7	Otras	613	8427	7,2742	100,000	1490	30042	4,9597	100,000

- El número de casos registrados (N) de la *tabla 23* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

- La **centralidad** de los datos del escenario 1 se localiza entre la clase 3 y 4 mientras que para el escenario 2, entre las clases 2 y 3. La clase modal considerando los accidentes a cielo abierto se corresponde con un tipo de causa previa PC=4 (22,52% de los datos). Para el caso de minería subterránea, la clase modal corresponde a un tipo de causa previa PC=2 (37,45% de los datos). Así pues, se verifican los resultados expuestos en el apartado 3.3 del presente trabajo.

- Las **distribuciónes** de los datos de la variable PC son para ambos casos asimétricas y positivas. Sin embargo, esta tendencia es más pronunciada en el escenario 2 que en el 1. Los picos de la primera distribución son más bien bajos, destacando las clases PC=4 (22,52% de los datos) y PC=5 (22,01%). La segunda distribución sí que presenta unos picos claros sobre las clases 2 (37,45%) y 3 (26,31%).

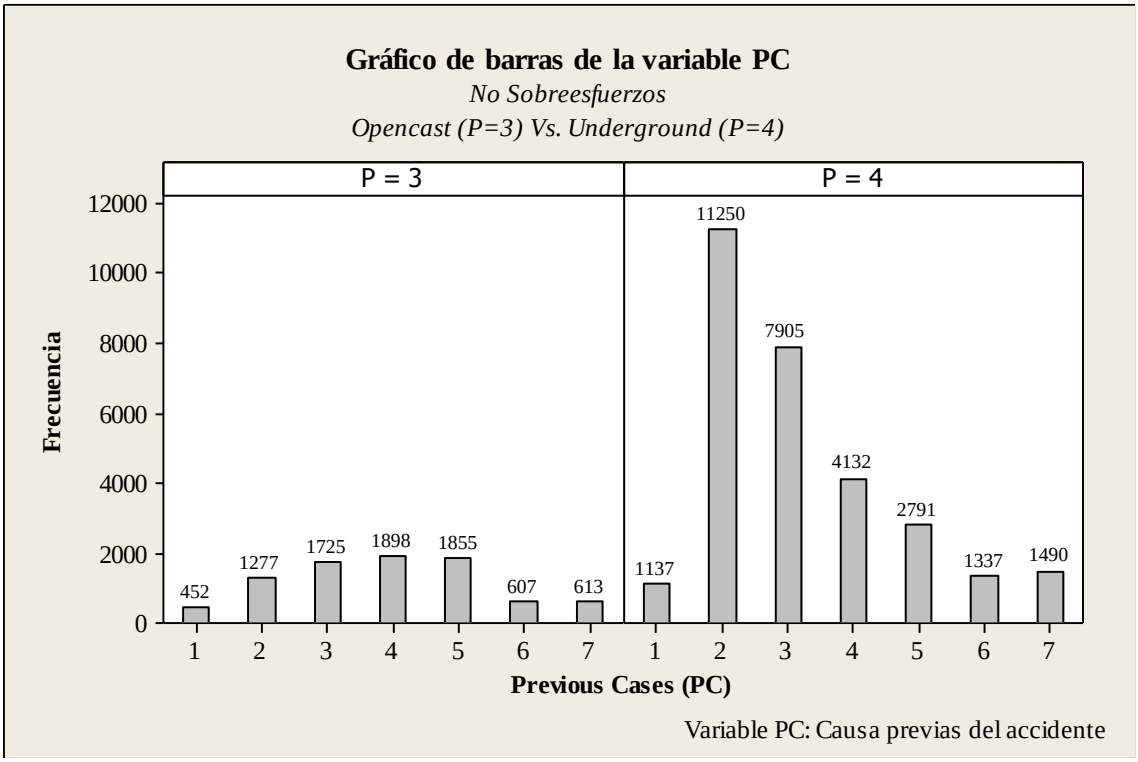

Figura 5: Gráfico de barras comparativo de la variable causas previas (PC) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

6) PA *(Physical Activity)*

Tabla 24: Tabla comparativa de frecuencias de la variable actividad física desarrollada (PA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

PA	Tipo de actividad física	Lugar accidente *(place)*							
		Cielo abierto (3_*opencast*)				Subterráneo (*4_underground*)			
		N	NAcum	%	% Acum	N	NAcum	%	% Acum
1	Operaciones con máquinas	1087	1087	12,8990	12,899	1065	1065	3,5450	3,545
2	Trabajos con herramientas manuales	1596	2683	18,9391	31,838	11256	12321	37,4675	41,013
3	Conducir/estar a bordo de un medio de transporte	755	3538	8,9593	40,797	905	13226	3,0124	44,025
4	Manipulación de objetos	1690	5128	20,0546	60,852	9547	22773	31,7788	75,804
5	Transporte manual	331	5459	3,9279	64,780	2283	25056	7,5994	83,403
6	Movimiento	2638	8097	31,3041	96,084	3910	28966	13,0151	96,418
7	Otras	330	8427	3,9160	100,000	1076	30042	3,5817	100,000

– El número de casos registrados (N) de la *tabla 24* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

- La **centralidad** de los datos de ambos escenarios se localizan entre las clases 3 y 4. La clase modal considerando los accidentes a cielo abierto se corresponde con un tipo de actividad física desarrollada de tipo PA=6 (31,30% de los datos). Para el caso de minería subterránea, la clase modal corresponde a un actividad física PA=2 (37,46% de los datos).

- Las **distribución** de los datos de la variable PC son opuestas. Para el primer caso (P=3) la asimetría es negativa mientras que para el segundo (P=4) resulta positiva. Los picos de la primera distribución son la clase 2, 4 y 6. Los de la segunda, son la clase 2, 4 y 5. Así pues, se verifican los resultados expuestos en el apartado 3.3 del presente trabajo.

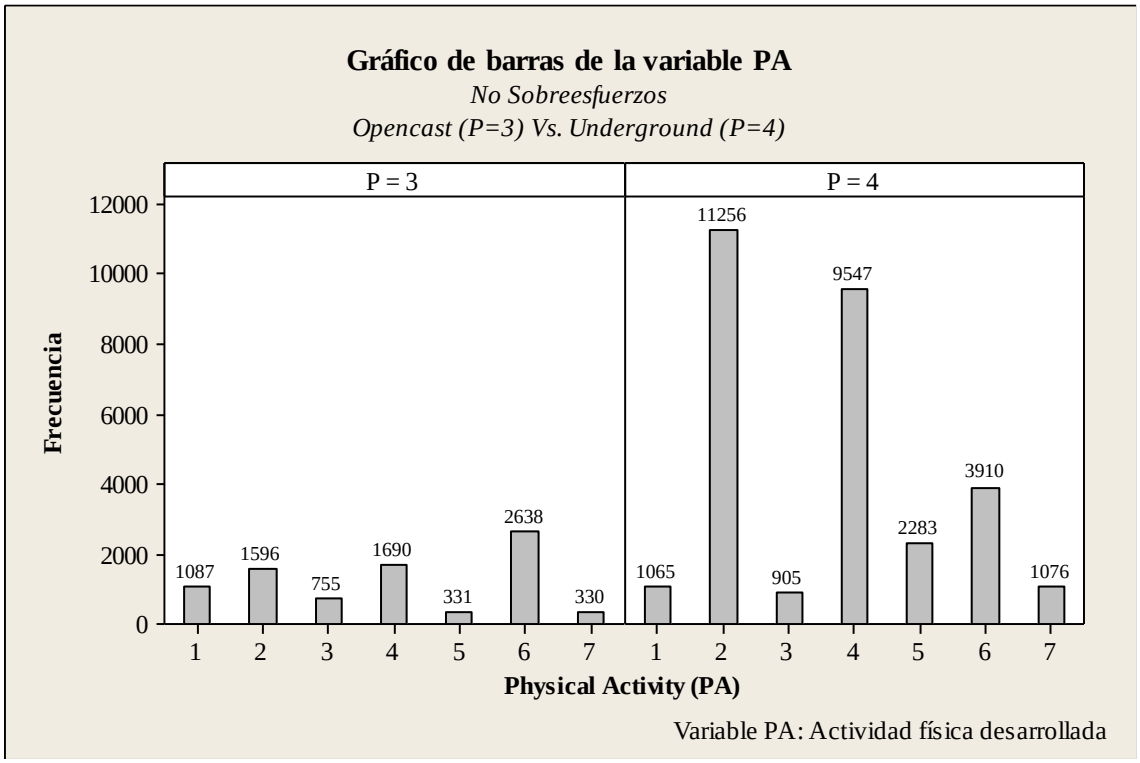

Figura 6: Gráfico de barras comparativo de la variable actividad física desarrollada (PA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

13) TA *(Type of Accident)*

Tabla 25: Tabla comparativa de frecuencias de la variable tipo de accidente (TA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

TA	Tipología accidente	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	Contacto eléctrico, fuego, contacto con sustancias peligrosas. Ahogamiento, quedar sepultado o envuelto.	343	343	4,0703	4,070	680	680	2,2635	2,263
2	Golpe contra un objeto inmóvil	2573	2916	30,5328	34,603	5093	5773	16,9529	19,216
3	Choque o golpe contra un objeto en movimiento o colisión	2450	5366	29,0732	63,676	13270	19043	44,1715	63,388
4	Contacto con objeto cortante, punzante, duro o rugoso	856	6222	10,1578	73,834	5277	24320	17,5654	80,953
5	Sobreesfuerzo físico, trauma psíquico, radiaciones, ruido, luz o presión	1270	7492	15,0706	88,905	3073	27393	10,2290	91,182
6	Otras	935	8427	11,0953	100,000	2649	30042	8,8177	100,000

- El número de casos registrados (N) de la *tabla 25* concuerda con los del escenarios 1 y 2 definidos en el apartado 2.2.

- La **centralidad** de los datos de ambos escenarios se localizan entre las clases 2 y 3. La clase modal considerando los accidentes a cielo abierto se corresponde con un tipo de accidente TA=2 (30,53% de los datos). Para el caso de minería subterránea, la clase modal corresponde a un tipo de accidente TA=3 (44,17% de los datos).

- Las **distribución** de los datos de la variable TA son asimétricas positivas. Los picos de la primera distribución son la clase 2 y 3; los de la segunda, únicamente destaca la clase 3.

- Así pues, se verifican los resultados expuestos en el apartado 3.3 del presente trabajo.

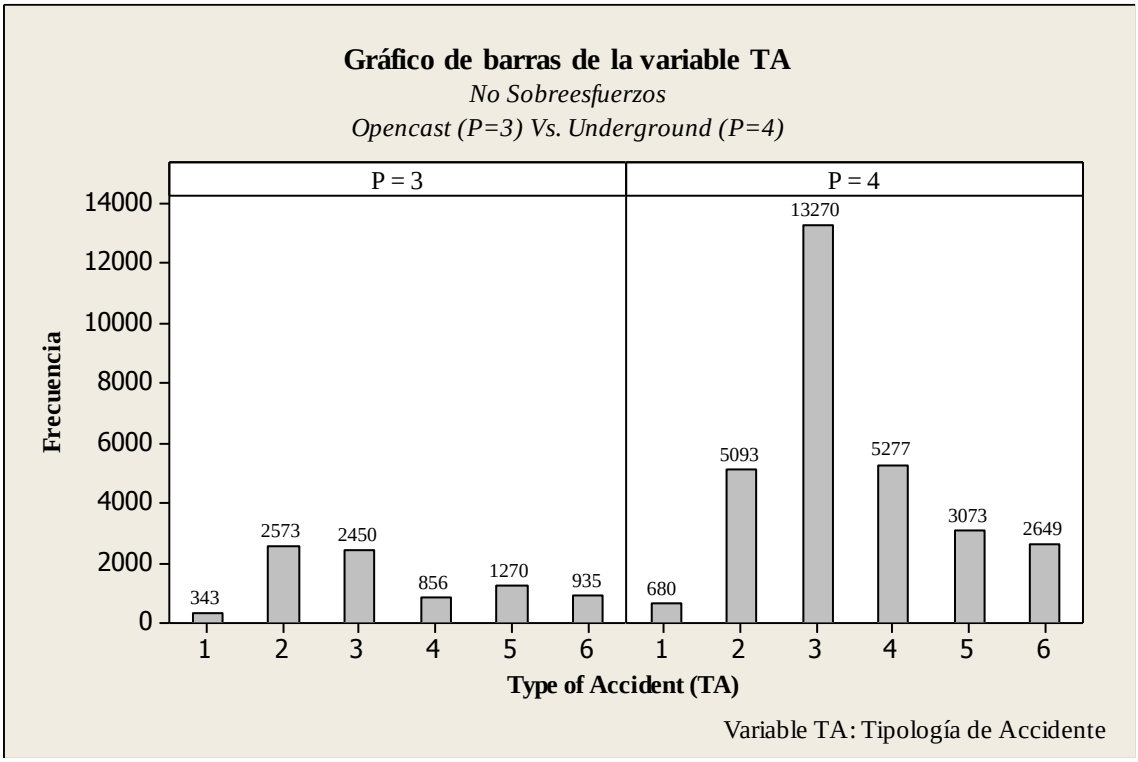

Figura 7: Gráfico de barras comparativo de la variable tipo de accidente (TA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

4.2.2. Accidentes registrados con sobreesfuerzos

Los accidentes con sobreesfuerzo son contemplados en los escenarios 3 y 4 del presente trabajo y, las variables a estudiar serán E, S, PC, PA, WH y TA.

2) E *(Experience)*

Tabla 26: Tabla comparativa de frecuencias de la variable experiencia (E) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

E	Meses	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	0-12	657	657	31,0932	31,093	820	820	12,3160	12,316
2	13-30	352	1009	16,6588	47,752	933	1753	14,0132	26,329
3	31-60	402	1411	19,0251	66,777	843	2596	12,6615	38,991
4	61-120	449	1860	21,2494	88,027	2292	4888	34,4248	73,415
5	121-180	119	1979	5,6318	93,658	642	5530	9,6425	83,058
6	181-240	76	2055	3,5968	97,255	787	6317	11,8204	94,878
7	240 o más	58	2113	2,7449	100,000	341	6658	5,1217	100,000

– El número de casos registrados (N) de la *tabla 26* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

– La **centralidad** de los datos para ambos emplazamientos es diferente. Para cielo abierto ésta se encuentra comprendida entre las clases 2 y 3 mientras que para minería subterránea se

localiza entre las clases 3 y 4. El tercer cuartil del primer caso corresponde a la clase 4 y evidencia el poco peso que aportan las tres últimas clases del gráfico de barras. Para el segundo caso, el tercer cuartil coincide con la categoría 5. La clase modal del primer escenario es la 1 mientras que para la el segundo es la 4.

– Acerca de la **distribución** de los datos, se observan asimetrías positivas en ambos emplazamientos. Los picos de la primera clase son más bien bajos, destacando el de la clase 1. En el escenario 4 el único pico que existe es el que corresponde a la clase 4.

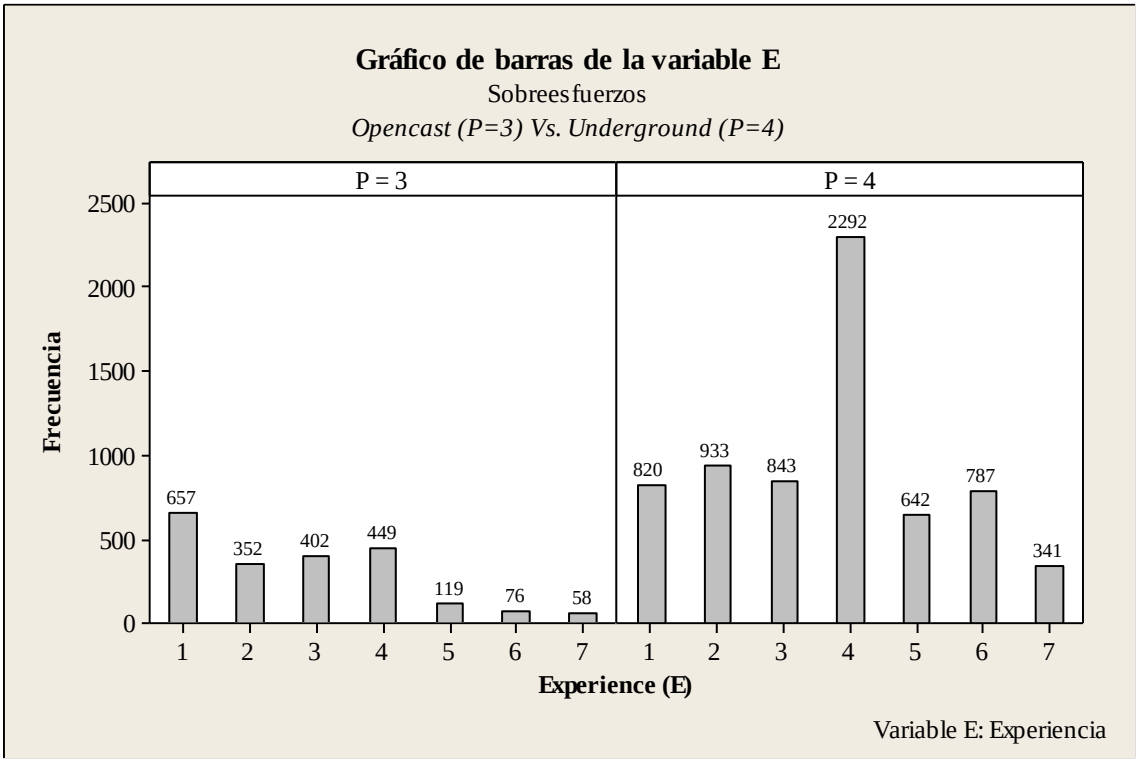

Figura 8: Gráfico de barras comparativo de la variable experiencia (E) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

3) S *(Size)*

Tabla 27: Tabla comparativa de frecuencias de la variable tamaño de empresa (S) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

S	Número de empleados	**Lugar accidente** *(place)*							
		Cielo abierto (*3_opencast*)				**Subterráneo** (*4_underground*)			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	0-9	534	534	25,2721	25,272	87	89	1,3067	1,307
2	10-19	582	1116	27,5438	52,816	78	165	1,1715	2,478
3	20-49	714	1830	33,7908	86,607	519	684	7,7951	10,273
4	50-99	164	1994	7,7615	94,368	651	1335	9,777	20,051
5	100-499	116	2110	5,4898	99,858	3364	4699	50,5257	70,577
6	500 o más	3	2113	0,1420	100,000	1959	6658	29,4233	100,000

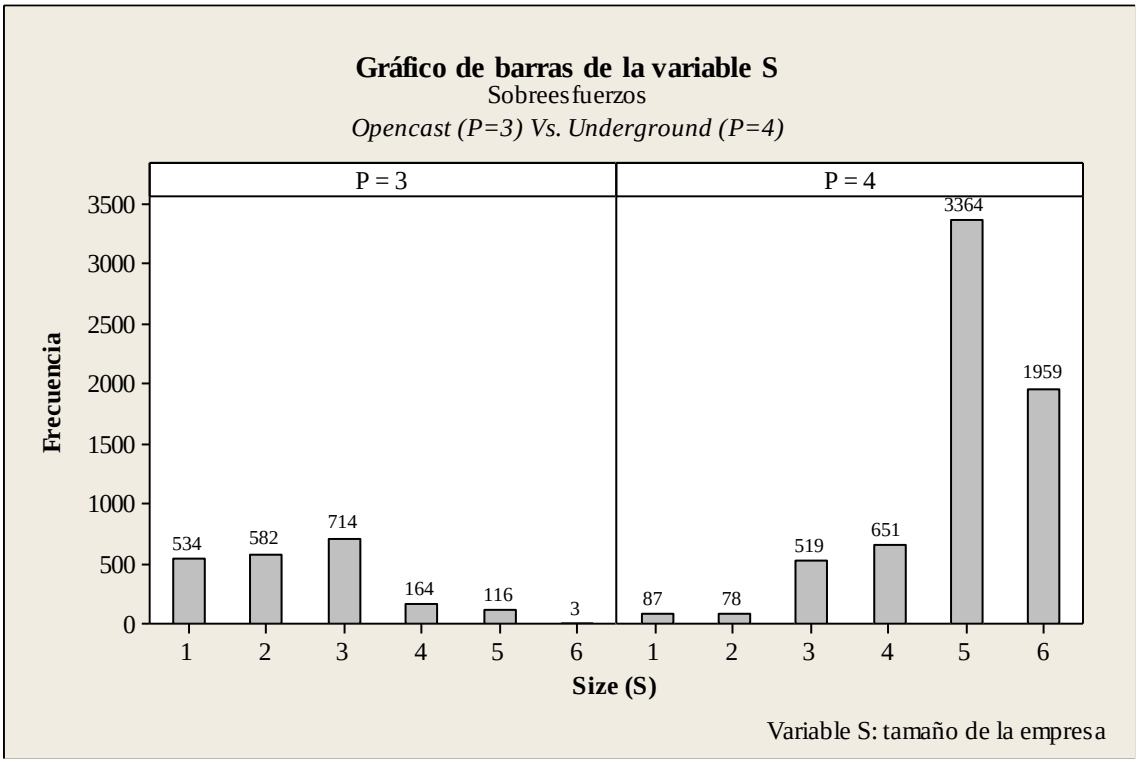

Figura 9: Gráfico de barras comparativo de la variable tamaño de empresa (S) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

El número de casos registrados (N) de la *tabla 27* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

– La **centralidad** de los datos es diferente para el tipo de emplazamiento analizado. Para el caso de cielo abierto, está se encuentra contenida entre las clases 1 y 2 mientras que para la minería subterránea, ésta se localiza entre las categorías 4 y 5. La clase modal del escenario 3 es la 3 y la del escenario 4 es la 5, representando el 33,79% y 50,53% de los datos, respectivamente. El tercer cuartil a cielo abierto se ubica en la clase 3 mientras que en minería subterránea corresponde a la clase 6.

– En cuanto a la **distribución** de los datos de la variable S, se observan comportamientos opuestos entre P=3 y P=4. No se aprecia ningún tipo de simetría y por tanto, la primera distribución corresponde a una asimetría positiva y la segunda a una negativa. Los picos del gráfico de barras cuando P=3 son la clase 1, 2 y 3. Para P=4, los picos de la distribución son las clases 5 y 6.

– Los resultados se corresponden con las hipótesis efectuadas en el apartado 3.3.

5) PC *(Previous Causes)*

Tabla 28: Tabla comparativa de frecuencias de la variable causas previas (PC) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

PC	Tipo de causa previa	Lugar accidente ***(place)***							
		Cielo abierto (3_*opencast*)				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	Problema eléctrico, explosión, fuego, desbordamiento, vuelco, escape, derrame, vaporización	0	0	0	0	0	0	0	0
2	Rotura, fractura, estallido, resbalón, caída, derrumbamiento de agente material	0	0	0	0	0	0	0	0
3	Pérdida de control (total o parcial) de control de máquinas	0	0	0	0	0	0	0	0
4	Caídas de personas	0	0	0	0	0	0	0	0
5	Movimiento del cuerpo sin esfuerzo físico	0	0	0	0	0	0	0	0
6	Movimiento del cuerpo con esfuerzo físico	2113	2113	100	100	6658	6658	100	100
7	Otras	0	0	0	100	0	0	0	100

- El número de casos registrados (N) de la *tabla 28* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

- En esta ocasión, no es necesario estudiar ni la centralidad ni distribución de los datos. Tampoco representar gráficamente los resultados habida cuenta que, tal y como se detalla en la *tabla 28* no existe otro tipo de causa previa que no sea cuando PC=6. Como se ha comentado anteriormente, esta codificación nos permite diferenciar entre accidentes con y sin sobreesfuerzo. Así pues, para que exista un sobreesfuerzo del sistema musculo esquelético, previamente debe existir una causa previa en la que el cuerpo realice esfuerzo físico.

- Se verifican los resultados expuestos en el apartado 3.3 sobre esta variable.

6) PA *(Physical Activity)*

Tabla 29: Tabla comparativa de frecuencias de la variable actividad física desarrollada (PA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

PA	Tipo de actividad física desarrollada**)**	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast)*				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	Operaciones con máquinas	146	146	6,9096	6,910	134	134	2,0126	2,013
2	Trabajos con herramientas manuales	364	510	17,2267	24,136	1359	1493	20,4115	22,242
3	Conducir/estar a bordo de un medio de transporte	112	622	5,3005	29,437	100	1593	1,5020	23,926
4	Manipulación de objetos	597	1219	28,2537	57,690	3229	4822	48,4980	72,424
5	Transporte manual	371	1610	18,5045	76,195	1199	6021	18,0084	90,433
6	Movimiento	495	2105	23,4264	99,621	554	6575	8,3208	98,753
7	Otras	8	2113	0,3786	100,000	83	6658	1,2466	100,000

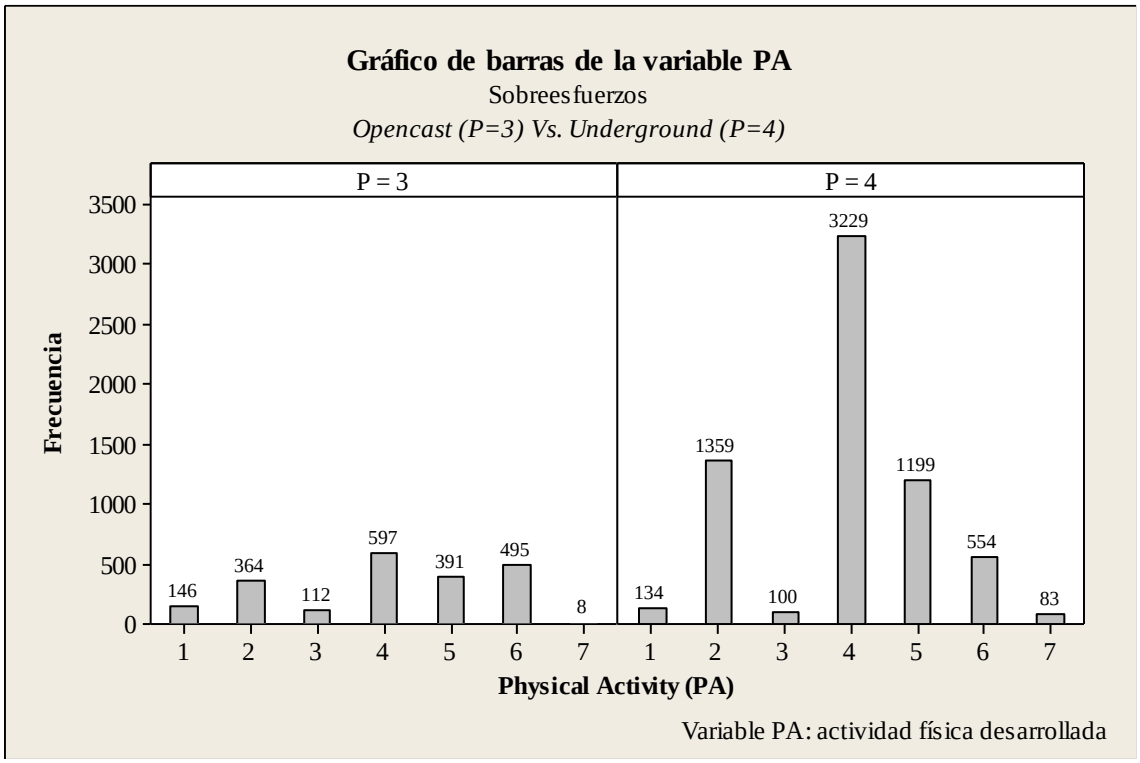

Figura 10: Gráfico de barras comparativo de la variable actividad física desarrollada (PA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

– El número de casos registrados (N) de la *tabla 29* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

– La **centralidad** de los datos de ambos escenarios se localizan entre las clases 3 y 4. La clase modal de ambos emplazamiento también coinciden con PA=4. Para el escenario 3 estos datos corresponden al 28,25% del total registrado mientras que para el escenario 4 este porcentaje aumenta hasta casi alcanzar la mitad de los datos registrados, es decir, el 48,50%.

– Las **distribución** de los datos de la variable PA es asimmétrica negativa en ambos casos. Los picos del escenario 3 son más bien bajos, aunque destaca la 4 y 6. Para el escenario 4, el único pico destacable es el de la clase 4.

11) WH *(Work Hours)*

Tabla 30: Tabla comparativa de frecuencias de la variable horas de jornada laboral completadas (WH) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

WH	Horas de jornada laboral trabajadas	**Lugar accidente** *(place)*							
		Cielo abierto (3_*opencast)*				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	(0, 1]	248	248	11,7369	11,737	679	679	10,1983	10,198
2	(1, 4]	1110	1358	52,5319	64,269	3633	4312	54,5659	64,764
3	(4, 8]	720	2078	34,0748	98,344	2279	6591	34,2295	98,994
4	(8, 10]	20	2098	0,9465	99,290	29	6620	0,4356	99,429
5	(10, 12]	9	2107	0,4259	99,176	17	6637	0,2553	99,685
6	(12 o más)	6	2113	0,2840	100,000	21	6658	0,3154	100,000

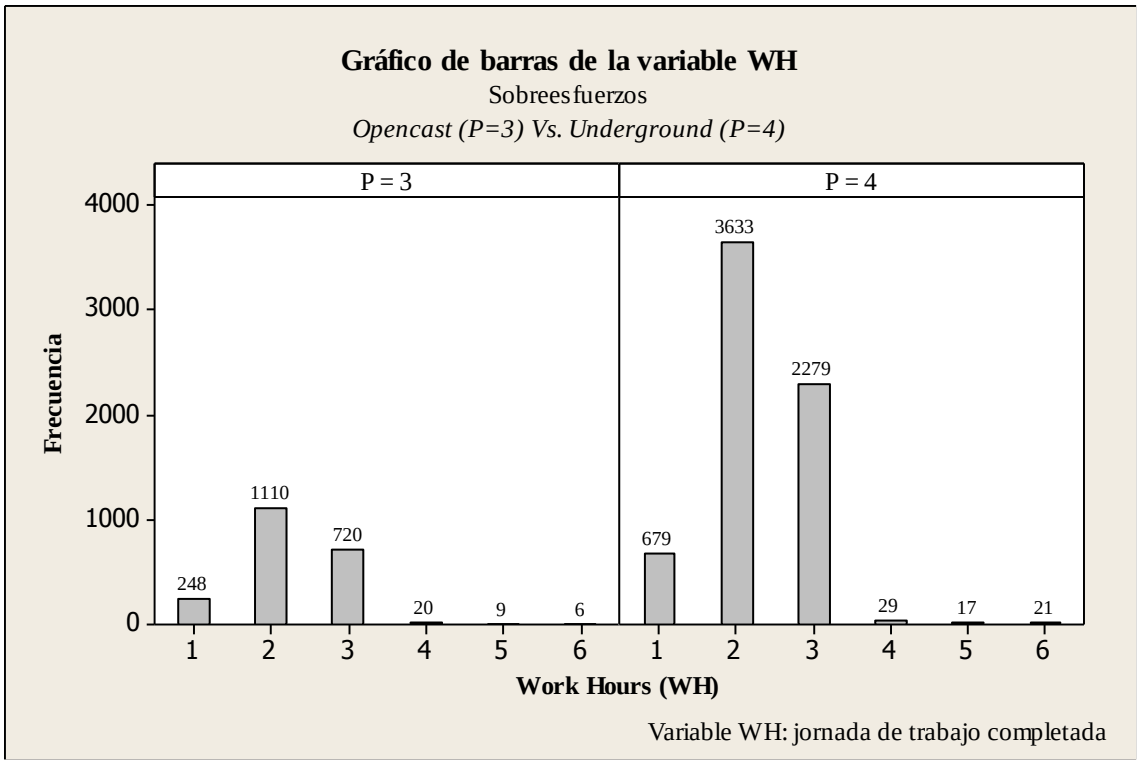

Figura 11: Gráfico de barras comparativo de la variable horas de jornada laboral completadas (WH) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

– El número de casos registrados (N) de la *tabla 30* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

– La **centralidad** de los datos de ambos escenarios se localizan entre las clases 1 y 2. La clase modal de ambos emplazamiento también coinciden cuando WH=2. Para el escenario 3 estos datos corresponden al 52,53% del total registrado mientras que para el escenario 4 este porcentaje representa el 54,57%. El tercer cuartil de ambos emplazamiento (P=3) se ubica en la clase 3, destacando así el poco peso que tienen el resto de clases sobre esta variable. De acuerdo con esto, se verifican los resultados expuestos en el apartado 3.3.

– Las **distribución** de los datos de la variable WH es asimmétrica negativa en ambos casos, muy similares y sobretodo con un grado de apilamiento elevado sobre la clase 2. Los de ambos emplazamientos corresponden a las clases 1,2 y 3.

– Se verifican de forma general los resultados expuestos en el apartado 3.3 sobre esta variable.

13) TA *(Type of Accident)*

Tabla 31: Tabla comparativa de frecuencias de la variable tipo de accidente (TA) entre accidentes laborales producidos a cielo abierto y subterráneo en el sector de la minería española durante el 2003 y 2013 (elaboración propia).

TA	Tipo de accidente	**Lugar accidente *(place)***							
		Cielo abierto (3_*opencast)*				**Subterráneo** *(4_underground)*			
		N	**NAcum**	**%**	**% Acum**	**N**	**NAcum**	**%**	**% Acum**
1	Contacto eléctrico, fuego, contacto con sustancias peligrosas. Ahogamiento, quedar sepultado o envuelto.	0	0	0	0	0	0	0	0
2	Golpe contra un objeto inmóvil	0	0	0	0	0	0	0	0
3	Choque o golpe contra un objeto en movimiento o colisión	0	0	0	0	0	0	0	0
4	Contacto con objeto cortante, punzante, duro o rugoso	0	0	0	0	0	0	0	0
5	Sobreesfuerzo físico, trauma psíquico, radiaciones, ruido, luz o presión	2113	2113	100	100	6658	6658	100	100
6	Otras	0	0	0	100	0	6658	0	100

– El número de casos registrados (N) de la *tabla 31* concuerda con los del escenarios 3 y 4 definidos en el apartado 2.2.

- En esta ocasión, al igual que cuando se analizaba la variable PC, estudiar ni la centralidad ni distribución de los datos. Tampoco representar gráficamente los resultados habida cuenta que, tal y como se detalla en la *tabla 31* no existe otro tipo de accidente que no sea cuando TA=5.

- Se verifican así los resultados expuestos sobre TA y PC en el apartado 3.3.

5. Conclusiones

A modo de conclusión, se confirma que:

- Se realizó un estudio de minería de datos en una muestra de 47.240 casos registrados durante el 2003 y el 2013 sobre accidentes laborales en minería del sector español y, diferenciando entre su emplazamiento y tipo de sobreesfuerzo.

- Se emplearon técnicas de clasificación, asociación y evaluación de pruebas de diversos algoritmos y, posteriormente se interpretaron los resultados.

- Se encontraron patrones de comportamiento singulares para cada uno de los escenarios propuestos.

- Se ha probado que uno de los factores fundamentales que mejor define la génesis de los accidentes laborales mineros del sector español, es la variable PC *(Previous Causes)*. Así queda reflejado en la clasificación de las siete variables más significativas y en los altos índices de confiabilidad de las reglas asociación obtenidas.

- La discriminación de accidentes en función del factor sobreesfuerzo da como resultado patrones de comportamiento bien diferentes.

- La discriminación de accidentes en función del emplazamiento no ha producido resultados tan concluyentes como los del factor sobreesfuerzo.

- Se ha comprobado que las técnicas de minería de datos son una importante y potencial fuente de información para multitud de ámbitos y disciplinas capaces de articular información supuestamente "escondida".

6. Bibliografía

Montero, J.M. (2007). *Estadística Descriptiva.* Madrid: Thomson.

Pérez, C., Santín, D. (2007). Minería de datos: técnicas y herramientas (1a ed.). Madrid: Thomson.

Sanmiquel, L., Freijo, M., Edo, J., Rossell, J.M. 2010. Analysis of work related accidents in the Spanish mining sector from 1982-2006. Journal of Safety Research. 41, 1-7.

Sanmiquel, L., Freijo, M., Rossell, J.M. 2012. Exploratory Analysis of Spanish Energetic Mining Accidents. International Journal of Ocupational Safety and Ergonomics. 18(2), 209-219.

Sanmiquel, L., Rossell, J.M., Freijo, M., Vintró, C. 2014. Influence of occupational safety management on the incidence rate of occupational accidents in the Spanish industrial and ornamental stone mining. WORK. 49, 307-314.

Sanmiquel, L., Rossell, J.M., Vintró, C. 2015. Study of Spanish mining accidents using data mining techniques. Safety Science. 75, 49-55.

Saari, J. 2005. Pequeñas y medianas empresas. Jornada técnica: La prevención de Riesgos Laborales en la PYME. Sevilla: Asepeyo, 10-20.

Williamson, A., Feyer, A. 1998. The causes of electrical fatalities at work. J. Saf. Res. 29 (3), 187-196.

Witten, I.H., Frank, E., Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Elvesier. USA.

.